Electronic Journal of Statistics
Vol. 10 (2016) 3548–3578
ISSN: 1935-7524
DOI: 10.1214/16-EJS1196

Simple approximate MAP inference for Dirichlet processes mixtures

Yordan P. Raykov[*]

Aston University
e-mail: yordan.raykov@gmail.com

Alexis Boukouvalas[*]

University of Manchester
e-mail: alexis.boukouvalas@manchester.ac.uk

and

Max A. Little

Aston University and MIT
e-mail: maxl@mit.edu

Abstract: The Dirichlet process mixture model (DPMM) is a ubiquitous, flexible Bayesian nonparametric statistical model. However, full probabilistic inference in this model is analytically intractable, so that computationally intensive techniques such as Gibbs sampling are required. As a result, DPMM-based methods, which have considerable potential, are restricted to applications in which computational resources and time for inference is plentiful. For example, they would not be practical for digital signal processing on embedded hardware, where computational resources are at a serious premium. Here, we develop a simplified yet statistically rigorous approximate maximum a-posteriori (MAP) inference algorithm for DPMMs. This algorithm is as simple as DP-means clustering, solves the MAP problem as well as Gibbs sampling, while requiring only a fraction of the computational effort.[†] Unlike related small variance asymptotics (SVA), our method is non-degenerate and so inherits the "rich get richer" property of the Dirichlet process. It also retains a non-degenerate closed-form likelihood which enables out-of-sample calculations and the use of standard tools such as cross-validation. We illustrate the benefits of our algorithm on a range of examples and contrast it to variational, SVA and sampling approaches from both a computational complexity perspective as well as in terms of clustering performance. We demonstrate the wide applicabiity of our approach by presenting an approximate MAP inference method for the infinite hidden Markov model whose performance contrasts favorably with a recently proposed hybrid SVA approach. Similarly, we show how our algorithm can applied to a semiparametric mixed-effects regression model where the random effects distribution is modelled using an infinite mixture model, as used in longitudinal progression modelling in population health science. Finally, we propose directions for future research on approximate MAP inference in Bayesian nonparametrics.

[*] These authors contributed equally to this work.

[†] For freely available code that implements the MAP-DP algorithm for Gaussian mixtures see http://www.maxlittle.net/.

MSC 2010 subject classifications: Primary 62F15; secondary 62G86.
Keywords and phrases: Bayesian nonparametrics, clustering, Gaussian mixture model.

Received May 2016.

1. Introduction

Bayesian nonparametric (BNP) models have been successfully applied to a wide range of domains but despite significant improvements in computational hardware, statistical inference in most BNP models remains infeasible in the context of large datasets, or for moderate-sized datasets where computational resources are limited. The flexibility gained by such models is paid for with severe decreases in computational efficiency, and this makes these models somewhat impractical. This is an important example of the emerging need for approaches to inference that simultaneously minimize both empirical risk and computational complexity (Bousquet and Bottou, 2008). Towards that end we study a simple, statistically rigorous and computationally efficient approach for the estimation of BNP models that significantly reduces the computational burden involved, while keeping most of the model properties intact. In this work, we concentrate on inference for the *Dirichlet process mixture model* (DPMM) and for the *infinite hidden Markov model* (iHMM) (Beal et al., 2002) but our arguments are more general and can be extended to many BNP models.

DPMMs are mixture models which use the *Dirichlet process* (DP) (Ferguson, 1973) as a prior over the mixing distribution of the model parameters. The data is modeled with a distribution with potentially infinitely many mixture components. The DP is an adaptation of the discrete Dirichlet distribution to the infinite, uncountable sample space. Where the Dirichlet distribution is formed over a continuous K-element sample space, if $K \to \infty$ we obtain the DP. A draw from a DP is itself a probability distribution. A DP is the Bayesian conjugate prior to the empirical probability distribution, much as the discrete Dirichlet distribution is conjugate to the categorical distribution. Hence, DPs have value in Bayesian probabilistic models because they are priors over completely general probability distributions. DPs can be also used as building blocks for more complex hierarchical models; an example being the the *hierarchical DP hidden Markov model* (HDP-HMM) for time series data, obtained by modeling the transition density in a standard HMM with a *hierarchical Dirichlet process* (HDP) (Teh et al., 2006).

An interesting property of DP-distributed functions is that they are discrete in the following sense: they are formed of an infinite, but countable mixture of Dirac delta functions. Since the Dirac has zero measure everywhere but at a single point, the support of the function is also a set of discrete points. This discreteness means that draws from such distributions have a non-zero probability of being repeats of previous draws. Furthermore, the more often a sample is repeated, the higher the probability of that sample being drawn again – an effect known as the “rich get richer” property (known as *preferential attachment* in the network science literature (Barabási and Albert, 1999)). This repetition, coupled with preferential attachment, leads to another valuable property of DPs: samples from DP-distributed densities have a strong *clustering* property whereby N draws can be partitioned into K representative draws, where $K \leq N$ and K is not fixed *a-priori*.

Inference in probabilistic models for which closed-form statistical estimation is intractable, is often performed using computationally demanding *Markov-chain Monte Carlo* (MCMC) techniques (Neal, 2000a; Teh et al., 2006; Van Gael et al., 2008), which generate samples from the distribution of the model parameters given the data. Despite the asymptotic convergence guarantees of MCMC, in practice MCMC often takes too long to converge and this can severely limit the range of applications. A popular alternative is to cast the inference problem as an optimization problem for which *variational Bayes* (VB) techniques can be used. Blei and Jordan (2004) first introduced VB inference for the DPMM, but their approach involves truncating the variational distribution of the joint DPMM posterior. Subsequently, collapsed variational methods (Teh et al., 2008) reduced the inevitable truncation error by working in a reduced-dimensional parameter space, but they are based on a sophisticated family of marginal likelihood bounds for which optimization is challenging. Streaming variational methods (Broderick et al., 2013a) obtain significant scaling by optimizing local variational bounds on batches of data visiting data points only once, but as a result they can easily become trapped in poor local fixed points. Similarly, stochastic variational methods (Chong et al., 2011) also allow for a single pass through the data, but sensitivity to initial conditions increases substantially. Alternatively, methods which learn memoized statistics of the data in a single pass (Hughes and Sudderth, 2013; Hughes et al., 2015) have recently shown significant promise.

Daumé (2007) describe a related approach for inference in DPMM based on a combinatorial search that is guaranteed to find the optimum for objective functions which have a specific computationally tractability property. As the DPMM complete data likelihood does not have this particular tractability property, their algorithm is only approximate for the DPMM, and this also makes it sample-order dependent. On the other hand, Dahl (2009) describe an algorithm that is guaranteed to find the global optimum in $N(N+1)$ computations, but only in the case of univariate product partition models with non-overlapping clusters. By contrast, our approach does not make any further assumptions beyond the model structure and being derived from the Gibbs sampler does not suffer from sample-order dependency. Wang and Dunson (2011) present another approach for fast inference in DPMMs which discards the exchangeability assumption of the data partitioning and instead assumes the data is in the correct ordering. Then a greedy, repeated "uniform resequencing" is proposed to maximize a pseudo-likelihood that approximates the DPMM complete data likelihood. This procedure does not have any guarantees for convergence even to a local optima. Zhang et al. (2014) extend the SUGS algorithm introducing a variational approximation of the cluster allocation probabilities. This allows replacement of the greedy allocation updates with updates of an approximation of the allocation distribution. However, this extension also lacks optimality guarantees and is mostly useful in streaming data applications.

Broderick et al. (2013b) propose a general approach to solving the MAP problem for a wide set of BNP models by forcing the spread of the likelihood of BNP models to zero. By making some additional simplifying assumptions, this approach reduces MCMC updates to a fast optimization algorithm that converges quickly to an approximate MAP solution. However, this *small variance asymptotic* (SVA) reasoning breaks many of the key properties of the underlying probabilistic model: SVA applied to the DPMM (Kulis and Jordan, 2012; Jiang et al., 2012) loses the rich-get-richer effect of the infinite clustering, as the prior term over the partition drops from the likelihood; and degeneracy in the likelihood forbids any kind of rigorous out-of-sample prediction and thus, for example, cross-validation. Roychowdhury et al. (2013) impose somewhat more flexible SVA assumptions to derive an optimization algorithm for inference in the *infinite hidden Markov model* (iHMM). Although this approach overcomes some of the drawbacks of SVA Broderick et al. (2013b), the algorithm departs from the assumptions of the underlying probabilistic graphical model. The method is shown to be efficient for clustering time dependent data, but essentially no longer has an underlying probabilistic model. Furthermore, Roychowdhury et al. (2013) demonstrate that there is more than one way of applying the SVA concept to a given probabilistic model, and therefore, under different choices of SVA assumptions, one obtains entirely different inference algorithms that find different structures in the data, even though the underlying probabilistic model remains the same. For example, HDP-means (Jiang et al., 2012) in the context of time series, and the alternative SVA approach of Roychowdhury et al. (2013) optimize different objective functions, even though they address inference for identical probabilistic models. To clarify this and other issues, we present a novel, unified exposition of the SVA approach in Section 4, highlighting some of its deficiencies and we show how these can be overcome using the non-degenerate MAP inference algorithms proposed in this paper.

In Section 2 we review the collapsed Gibbs sampler for DPMMs and in Section 3 we show how the collapsed Gibbs sampler may be exploited to produce simplified MAP inference algorithms for DPMMs. As with DP-means it provides only point estimates of the joint posterior. However, while DP-means follows the close relationship between K-means and the (finite) *Gaussian mixture model* (GMM) to derive a "nonparametric K-means", we exploit the concept of *iterated conditional modes* (ICM) (Kittler and Föglein, 1984). Experiments on both synthetic and real-world datasets are used to contrast the MAP-DP, collapsed Gibbs, DP-means and variational DP approaches in Section 5. In Section 6 we demonstrate how the MAP DPMM approach can be extended to the iHMM and contrast it to the hybrid SVA approach of Roychowdhury et al. (2013) in a simulation study. Finally, we demonstrate an application of our new algorithm to a hierarchical model of longitudinal health data in Section 7 and conclude with a discussion of future directions for this MAP approach for BNP models in Section 8.

2. Collapsed Gibbs sampling for Dirichlet process mixtures

The DPMM is arguably the most popular Bayesian nonparametric model which extends finite mixture models to the infinite setting by use of the DP prior. In this work we will restrict ourselves to mixture models with exponential family distribution data likelihoods. We will denote by $\mathbf{X}$ the full data matrix formed of the observed data points $\mathbf{x}_i$ which are D-dimensional vectors $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d}, \ldots, x_{i,D})$, N_0 is the concentration parameter of the DP prior and G_0 is its base measure. The DPMM is then often written as:

$$\begin{array}{rcl} G & \sim & \mathrm{DP}\left(N_0, G_0\right) \\ \vartheta_i \left| G \right. & \overset{i.i.d.}{\sim} & G, \;\; i = 1, \ldots, N \\ \mathbf{x}_i \left| \vartheta_i \right. & \sim & F\left(\mathbf{x}_i; \vartheta_i\right), \;\; i = 1, \ldots, N \end{array} \tag{2.1}$$

where G is a mixing distribution drawn from a DP; ϑ are the atoms of G which take repeated values and F is the distribution of each data point given its atom. We can also write the mixing distribution G in terms of mixture weights π and the distrinct values taken from ϑdenoted with θ, $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}$ and $\mathbf{x}_i \sim \sum_{k=1}^{\infty} \pi_k F\left(\theta_k\right)$, where $\delta\left(\cdot\right)$ denotes the Dirac delta function. The probability of the data follows an infinite mixture distribution and because this likelihood is not available in closed form, a Gibbs sampling procedure is not tractable. A widely used approach to overcome this issue is to collapse the mixture weights and model the data in terms of the cluster indicator variables $z_1, \ldots, z_N$:

$$\begin{array}{rcl} \left(z_1, \ldots, z_N\right) & \sim & \mathrm{CRP}\left(N_0, N\right) \\ \theta_1, \ldots, \theta_K \left| z \right. & \overset{i.i.d.}{\sim} & G_0 \\ \mathbf{x}_i \left| z, \theta \right. & \sim & F\left(\mathbf{x}_i; \theta_{z_i}\right), \;\; i = 1, \ldots, N \end{array} \tag{2.2}$$

where to simplify notation we denote $z = (z_1, \ldots, z_N)$ and $\theta = (\theta_1, \ldots, \theta_K)$; CRP stands for the *Chinese restaurant process* which is a discrete stochastic process over the space of partitions, or equivalently a probability distribution over cluster indicator variables. It is strictly defined by the integer N (number of observed data points) and a positive, real *concentration* parameter N_0. A draw from a CRP has probability:

$$p\left(z_1, \ldots, z_N\right) = \frac{\Gamma\left(N_0\right)}{\Gamma\left(N + N_0\right)} N_0^K \prod_{k=1}^{K} \Gamma\left(N_k\right) \tag{2.3}$$

with indicators $z_1, \ldots, z_N \in \{1, \ldots, K\}$, where K is the unknown number of items and $N_k = \left|\{i : z_i = k\}\right|$ is the number of indicators taking value k, with $\sum_{k=1}^{K} N_k = N$. For any finite N we will have $K \leq N$ and usually K will be much smaller than N, so the CRP returns a *partition* of N elements into some smaller number of groups K. The probability over indicators is constructed in a sequential manner using the following conditional probability:

$$p\left(z_{n+1}=k\left|z_{1},\ldots,z_{n}\right.\right)=\begin{cases}\frac{N_{k}}{N_{0}+n} & \text{if } k=1,\ldots,K\\ \frac{N_{0}}{N_{0}+n} & \text{otherwise}\end{cases} \tag{2.4}$$

By increasing the value of n from 1 to N and using the corresponding conditional probabilities, we obtain the joint distribution over indicators from Equation (2.3), $p\left(z_{1},\ldots,z_{N}\right)=p\left(z_{N}\left|z_{1},\ldots,z_{N-1}\right.\right)p\left(z_{N-1}\left|z_{1},\ldots,z_{N-2}\right.\right)\times\cdots\times p\left(z_{2}\left|z_{1}\right.\right)$.

The probability density function of $\mathbf{x}_{i}\sim F\left(\mathbf{x}_{i};\theta_{z_{i}}\right)$ associated with the component indicated by the value of z_{i}, is an exponential family distribution:

$$p\left(\mathbf{x}_{i}\left|\theta_{z_{i}}\right.\right)=\exp\left(\left\langle \boldsymbol{g}\left(\mathbf{x}_{i}\right),\theta_{z_{i}}\right\rangle -\psi\left(\theta_{z_{i}}\right)-h\left(\mathbf{x}_{i}\right)\right) \tag{2.5}$$

where $\boldsymbol{g}\left(.\right)$ is the sufficient statistic function, $\psi(\theta_{z_{i}})=\log\int\exp(\langle\mathbf{x}_{i},\theta_{z_{i}}\rangle-h(\mathbf{x}_{i}))d\mathbf{x}_{i}$ is the log partition function and $h\left(\mathbf{x}_{i}\right)$ the base measure of the distribution. An important property of exponential family distributions is that the conjugate prior over the natural parameters $\theta_{k}\sim G_{0}$ exists and can be obtained in closed form:

$$p\left(\theta\left|z,\boldsymbol{\tau},\eta\right.\right)=\exp\left(\left\langle \theta,\boldsymbol{\tau}\right\rangle -\eta\psi\left(\theta\right)-\psi_{0}\left(\boldsymbol{\tau},\eta\right)\right) \tag{2.6}$$

where $(\boldsymbol{\tau},\eta)$ are the prior hyperparameters of the base measure G_{0}, ψ_{0} is base measure of the parameter distribution. From Bayesian conjugacy, the posterior $p\left(\theta_{k}\left|\mathbf{X},\boldsymbol{\tau}_{k},\eta_{k}\right.\right)$ will take the same form as the prior where the prior hyperparameters $\boldsymbol{\tau}$ and η will be updated to $\boldsymbol{\tau}_{k}=\boldsymbol{\tau}+\sum_{j:z_{j}=k}\boldsymbol{g}\left(\mathbf{x}_{j}\right)$ and $\eta_{k}=\eta+N_{k}$.

Inference can be accomplished via *collapsed Gibbs sampling*, presented as Algorithm 3 in Neal (2000b). This MCMC algorithm iteratively samples each component indicator z_{i} for $i=1,\ldots,N$, conditional on all others, until convergence:

$$p\left(z_{i}=k\left|\mathbf{x}_{i},z_{-i}\right.\right)\propto\begin{cases}N_{k,-i}p\left(\mathbf{x}_{i}\left|\boldsymbol{\tau}_{k,-i},\eta_{k,-i}\right.\right) & \text{for existing } k\\ N_{0}p\left(\mathbf{x}_{i}\left|\boldsymbol{\tau},\eta\right.\right) & \text{for some new } k=K+1\end{cases} \tag{2.7}$$

where the subscript $-i$ denotes the removal of point i from consideration, $z_{-i}=\left\{z_{1},\ldots,z_{i-1},z_{i+1},\ldots,z_{N}\right\}$ and $p\left(\mathbf{x}_{i}\left|\boldsymbol{\tau},\eta\right.\right)$ is the posterior predictive density of point i obtained after integrating out the cluster parameters, $p\left(\mathbf{x}_{i}\left|\boldsymbol{\tau},\eta\right.\right)=\int p\left(\mathbf{x}_{i}\left|\theta\right.\right)p\left(\theta\left|\boldsymbol{\tau},\eta\right.\right)d\theta$. Examples of posterior predictive densities for different exponential family likelihoods are presented in Appendix B.

3. Introducing MAP-DP: A novel approximate MAP algorithm for collapsed DPMMs

In this section we propose a novel DPMM inference algorithm based on iteratively updating the cluster indicators with the values that maximize their posterior (MAP values). The cluster parameters are integrated out. This algorithm can be also seen as an an "exact" version of the *maximization-expectation* (M-E) algorithm presented in Welling and Kurihara (2006). It is exact in the following sense: while the M-E algorithm is a kind of VB and as with VB makes a factorization assumption which departs from the underlying probabilistic model purely for computational simplicity and tractability purposes, our algorithm is derived directly from the Gibbs sampler for the probabilistic model. Therefore, our algorithm does not introduce or require this simplifying factorization assumption. In Section 3.1 we describe how parameter inference for the DPMM is accomplished and in Section 3.2 we consider out-of-sample prediction.

3.1. Inference

As a starting point, we consider the DPMM introduced in Section 2. In our algorithm, we iterate through each of the cluster indicators z_i and update them with their respective MAP values. For each observation $\mathbf{x}_i$, we compute the negative log probability for each existing cluster k and for a new cluster $K+1$:

$$q_{i,k} = -\log p\left(\mathbf{x}_i | z_{-i}, \mathbf{X}_{-i}, z_i = k, \boldsymbol{\tau}_{k,-i}, \eta_{k,-i}\right) \tag{3.1}$$

$$q_{i,K+1} = -\log p\left(\mathbf{x}_i | \boldsymbol{\tau}, \eta\right) \tag{3.2}$$

where terms independent of k may be omitted as they do not change with k. For each observation $\mathbf{x}_i$ we compute the above $K+1$-dimensional vector $\boldsymbol{q}_i$ and select the cluster number according to the following:

$$z_i = \underset{k \in \{1,\dots,K,K+1\}}{\arg\min} \left[q_{i,k} - \log N_{k,-i}\right]$$

where $N_{k,-i}$ is the number of data points assigned to cluster k, excluding data point $\mathbf{x}_i$ and, for notational convenience, we define $N_{K+1,-i} \equiv N_0$.

The algorithm proceeds to the next observation $\mathbf{x}_{i+1}$ by updating the cluster component statistics to reflect the new value of the cluster assignment z_i and remove the effect of data point $\mathbf{x}_{i+1}$. To check convergence of the algorithm we compute the negative log of the complete data likelihood:

$$p\left(\mathbf{x}, z | N_0\right) = \left(\prod_{i=1}^{N} \prod_{k=1}^{K} p\left(\mathbf{x}_i | z_i\right)^{\delta(z_i,k)}\right) p\left(z_1, \dots, z_N\right) \tag{3.3}$$

where $\delta\left(z_i, k\right)$ is the Kronecker delta and $p\left(z_1, \dots z_N\right)$ is the CRP partition function (Pitman, 1995) given in Equation (2.3). We show in Algorithm 1 all the steps involved in approximately maximizing this complete data likelihood.

It is worth pointing out that unlike MCMC approaches, MAP-DP does not increase the negative log of the complete data likelihood at each step and as a result is guaranteed to converge to a fixed point. Where the convergence reached by MCMC sampling is convergence in distribution to the stationary posterior measure, the convergence of MAP-DP is only to local maxima of the posterior measure and so it is much quicker to reach. The main disadvantages with this are that the solution at convergence is only guaranteed to be a local maximum and that information about the whole distribution of the posterior is lost. Multiple restarts using random permutations of the data can be used to overcome poor local maximum. With MAP-DP it is possible to learn all model hyperparameters as we discuss in Appendix A and this is a strong advantage over the fast SVA approaches.

```
Input: $\mathbf{x}_1, \dots, \mathbf{x}_N$: data; $N_0 > 0$: concentration parameter, $\epsilon > 0$: convergence
    threshold; $(\boldsymbol{\tau}, \eta)$: cluster prior parameters; $\psi_0(.)$: prior log partition function;
    $\boldsymbol{g}(.)$: sufficient statistic function.
Output: $z_1, \dots, z_N$: cluster assignments, $K$: number of clusters.
$K = 1, z_i = 1$, for all $i \in 1, \dots, N$;
$E_{\text{new}} = \infty$;
repeat
    $E_{\text{old}} = E_{\text{new}}$;
    for $i \in 1, \dots, N$ do
        for $k \in 1, \dots, K$ do
            $q_{i,k} =$
            $\psi_0\left(\boldsymbol{\tau} + \sum_{j:z_j=k, j\neq i} g\left(\mathbf{x}_j\right), \eta + N_{k,-i}\right) - \psi_0\left(\boldsymbol{\tau} + \sum_{j:z_j=k} g\left(\mathbf{x}_j\right), \eta + N_k\right)$
        end
        $q_{i,K+1} = \psi_0\left(\boldsymbol{\tau}, \eta\right) - \psi_0\left(\boldsymbol{\tau} + \boldsymbol{g}(\mathbf{x}_i), \eta + 1\right)$;
        $z_i = \arg\min_{k \in 1,\dots,K,K+1}\left[q_{i,k} - \log N_{k,-i}\right]$;
        if $z_i = K + 1$ then
            $K = K + 1$;
        end
    end
    $E_{\text{new}} = \sum_{k=1}^{K}\sum_{i:z_i=k} q_{i,k} - K\log\left(N_0\right) - \sum_{k=1}^{K}\log\Gamma\left(N_k\right)$;
until $E_{\text{old}} - E_{\text{new}} < \epsilon$;
```

Algorithm 1: MAP-DP: Exponential Families

3.2. Out-of-sample prediction

To compute the out-of-sample likelihood for a new observation $\mathbf{x}_{N+1}$ we consider two approaches that differ in how the indicator z_{N+1} is treated:

1. *Mixture predictive density.* The unknown indicator z_{N+1} can be integrated out resulting in a mixture density:

$$p\left(\mathbf{x}_{N+1}|N_0, z, \mathbf{X}\right) = \sum_{k=1}^{K+1} p\left(z_{N+1} = k|N_0, z, \mathbf{X}\right) p\left(\mathbf{x}_{N+1}|z, \mathbf{X}, z_{N+1} = k\right) \qquad (3.4)$$

   The assignment probability $p\left(z_{N+1} = k|z, N_0, \mathbf{X}\right)$ is $\frac{N_k}{N_0+N}$ for an existing cluster and $\frac{N_0}{N_0+N}$ for a new cluster. The second term corresponds to the predictive distribution of point $N+1$ according to the predictive densities $p\left(\mathbf{x}_{N+1}|z, \mathbf{X}, \boldsymbol{\tau}_k, \eta_k, z_{N+1} = k\right)$ and $p\left(\mathbf{x}_{N+1}|\boldsymbol{\tau}, \eta, z_{N+1} = K+1\right)$ for an existing and new cluster respectively.

2. *MAP cluster assignment.* We can also use a point estimate for z_{N+1} by picking the minimum negative log posterior of the indicator $p\left(z_{N+1}|\mathbf{x}_{N+1}, N_0, z, \mathbf{X}\right)$, equivalently:

$$z_{N+1}^{\text{MAP}} = \underset{k\in\{1,\dots,K,K+1\}}{\arg\min}\left[-\log p\left(\mathbf{x}_{N+1}|z, \mathbf{X}, z_{N+1} = k\right) - \log p\left(z_{N+1} = k|N_0, z, \mathbf{X}\right)\right] \qquad (3.5)$$

where $p\left(\mathbf{x}_{N+1}|z,\mathbf{X},z_{N+1}=k\right)$ and $p\left(z_{N+1}=k|N_0,z,\mathbf{X}\right)$ are exactly as above. This approach is useful for clustering applications when we are interested in estimating z_{N+1}^{MAP} explicitly. Once the MAP assignment for point $N+1$ is updated, we estimate the probability of $\mathbf{x}_{N+1}$ given that it belongs to the component pointed by z_{N+1}^{MAP}, $p\left(\mathbf{x}_{N+1}|z,\mathbf{X},z_{N+1}^{\text{MAP}}\right)$.

The first (marginalization) approach is used in Blei and Jordan (2004) and is more "robust" as it incorporates the probability of all cluster components while the second (modal) approach can be useful in cases where only a point cluster assignment is needed. Integrating over variable z_{N+1} for more robust estimation of x_{N+1} is an example of the well studied process known as *Rao-Blackwellization* (Blackwell, 1947) which is often used in Bayesian inference for improving the quality of statistical estimation of uncertainty.

Even when using the first approach however, the mixture density is still computed assuming point assignments for the training data $z_1,\ldots,z_N$. Therefore the predictive density obtained using MAP-DP will be comparable to the one obtained using Gibbs sampler inference only when the sufficient statistics $N_1,\ldots N_K$ of the categorical likelihood for the assignment variables estimated from a Gibbs chain are similar to the ones estimated from the modal estimates for $z_1,\ldots,z_N$. Empirically, we have observed this often to be the case. Furthermore, we have noticed that the predictive density for popular (with a lot of points) cluster components tend to be well approximated by MAP-DP where the effect of the smaller cluster components diminishes when using only modal estimates for z. Note that the DPMM usually models data with a lot of inconsistent small spurious components (Miller and Harrison, 2013), those and any consistent components with small effect are likely to be ignored when using MAP-DP as we later show in Section 5.1. To summarize, using only modal estimates for the cluster assignments we are likely to infer correctly only larger components which have a large effect on the model likelihood and which will also affect the estimated predictive density accordingly.

4. Another look at small variance asymptotics (SVA)

The novel MAP-DP algorithm presented here has the "flavor" of an SVA-like algorithm, but there are critical differences and advantages, which we discuss in detail in this section.

Firstly, some background to the SVA approach is required. There exists a well known connection between the *expectation-maximization* (E-M) algorithm for the finite *Gaussian Mixture Model* (GMM) and K-means. That is, by assuming a GMM with equal variance, spherical (diagonal) component covariance matrices, we can obtain K-means from the E-M algorithm for the corresponding GMM by shrinking the component variances in each dimension to 0. This approach is more recently referred to as the *small variance asymptotic* (SVA) derivation of the K-means algorithm (Bishop, 2006, page 423).

Using *Bregman divergences* $D_{\phi}(\cdot)$ (see Appendix C), Banerjee et al. (2005) has extended the SVA reasoning to any exponential family finite mixtures and K-means like clustering procedures can be derived. Banerjee et al. (2005) showed that the likelihood of point $\mathbf{x}_i$ from component k given the component parameter θ_k and the posterior of parameter θ_k given its posterior hyperparameters can be rewritten using Bregman divergences as:

$$\begin{aligned} p\left(\mathbf{x}_i\left|\theta_k\right.\right) = & \exp\left(-D_{\phi}\left(\mathbf{x}_i, \boldsymbol{\mu}_k\right) f_{\phi}\left(\mathbf{x}_i\right)\right) \\ p\left(\theta_k\left|\boldsymbol{\tau}_k, \eta_k\right.\right) = & \exp\left(-\eta D_{\phi}\left(\frac{\boldsymbol{\tau}_k}{\eta_k}, \boldsymbol{\mu}_k\right)\right) g_{\phi}\left(\boldsymbol{\tau}_k, \eta_k\right) \end{aligned} \tag{4.1}$$

where ϕ is the Legendre-conjugate function of ψ, $f_{\phi}\left(\mathbf{x}_i\right) = \exp\left(\phi\left(\mathbf{x}_i\right) - h\left(\mathbf{x}_i\right)\right)$ and $g_{\phi}\left(\boldsymbol{\tau}, \eta\right) = \exp\left(\eta\phi\left(\theta\right) - \psi_0\left(\boldsymbol{\tau}, \eta\right)\right)$ with $h(\cdot)$ and $\psi_0(\cdot)$ denoting the base measure of the corresponding distributions and $\boldsymbol{\mu}$ is the expectation parameter satisfying $\boldsymbol{\mu} = \nabla\psi(\theta)$. Kulis and Jordan (2012) extended this more compact form to the nonparametric DPMM and with some further assumptions derived a nonparametric K-means like algorithm that we now review in detail. Consider the DPMM above (with non-integrated component parameters), but with a scaled exponential family likelihood $\tilde{F}\left(\tilde{\theta}\right)$ that is parameterized by a scaled natural parameter $\tilde{\theta} = \xi\theta$ and log-partition function $\tilde{\psi}\left(\tilde{\theta}\right) = \xi\psi\left(\tilde{\theta}/\xi\right)$ for some $\xi > 0$. Further assume that the prior parameters of the natural parameter are also scaled appropriately, such that $\tilde{\boldsymbol{\tau}} = \frac{\boldsymbol{\tau}}{\xi}$ and $\tilde{\eta} = \frac{\eta}{\xi}$. It is then straightforward to see that the conjugate prior of $\tilde{\psi}$ will be also scaled and so $\tilde{\phi} = \xi\phi$. Then Jiang et al. (2012) have shown that $\tilde{F}\left(\tilde{\theta}\right)$ has the same mean as $F(\theta)$, but scaled covariance, $\operatorname{cov}\left(\tilde{\theta}\right) = \operatorname{cov}(\theta)/\xi$. Let us also assume that N_0 is a function of ξ, η and τ, taking the form:

$$N_0 = \left(g_{\tilde{\phi}}\left(\frac{\boldsymbol{\tau}}{\xi}, \frac{\eta}{\xi}\right)\left(\frac{2\pi}{\xi+\eta}\right)^{D/2}\xi^{D}\right)^{-1}\exp\left(-\xi\lambda\right) \tag{4.2}$$

for some free parameter λ that will replace the concentration parameter in the new formulation; D denotes dimension of the data and is unrelated to $D_{\phi}(\cdot)$. Then we can write the scaled exponential family DPMM as $\xi \to \infty$ and as a consequence $\operatorname{cov}\left[\tilde{\theta}\right] \to 0$. Following Jiang et al. (2012) we can write out the Gibbs sampler probabilities in terms of $D_{\phi}(\cdot)$ (see Appendix C) after canceling out $f_{\tilde{\phi}}\left(\mathbf{x}_i\right)$ terms from all probabilities:

$$\begin{aligned} p\left(z_i = k\left|z_{-i}, \mathbf{x}_i, \xi, \boldsymbol{\mu}\right.\right) = & \frac{N_{k,-i}\exp\left(-\xi D_{\phi}\left(\mathbf{x}_i, \boldsymbol{\mu}_k\right)\right)}{C_{\mathbf{x}_i}\exp\left(-\xi\lambda\right) + \sum_{j=1}^{K} N_j\exp\left(-\xi D_{\phi}\left(\mathbf{x}_i, \boldsymbol{\mu}_j\right)\right)} \\ p\left(z_i = K+1\left|z_{-i}, \mathbf{x}_i, \xi, \boldsymbol{\mu}\right.\right) = & \frac{C_{\mathbf{x}_i}\exp\left(-\xi\lambda\right)}{C_{\mathbf{x}_i}\exp\left(-\xi\lambda\right) + \sum_{j=1}^{K} N_j\exp\left(-\xi D_{\phi}\left(\mathbf{x}_i, \boldsymbol{\mu}_j\right)\right)} \end{aligned}$$

where $C_{\mathbf{x}_i}$ approaches a positive, finite constant for a given $\mathbf{x}_i$ as $\xi \to \infty$ and we have used the fact that for a Bregman divergence, $D_{\xi\phi}\left(\centerdot, \centerdot\right) = \xi D_{\phi}\left(\centerdot, \centerdot\right)$. Now, both of the above probabilities will become binary (take on the values 0 or 1) as $\xi \to \infty$ and so all $K+1$ values will be increasingly dominated by the smallest value of $\{D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_1), D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_2), \ldots, D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_K), \lambda\}$. That is, the data point $\mathbf{x}_i$ will be assigned to the nearest cluster with Bregman divergence at most λ. If the closest mean has a divergence greater then λ, we create a new cluster containing only $\mathbf{x}_i$.

The posterior distribution over the cluster parameters for some component k is concentrated around the sample mean of points assigned to that component $\frac{1}{N_k}\sum_{i=1}^{N_k}\mathbf{x}_i$ as $\xi \to \infty$, so we update the cluster means with the sample mean of data points in each cluster, as with the corresponding parameter update step in K-means. The resulting algorithm approximately minimizes the following objective function over (z, μ):

$$\sum_{k=1}^{K}\sum_{i:z_i=k} D_\phi(\mathbf{x}_i, \boldsymbol{\mu}_k) + \lambda K \tag{4.3}$$

Similar objective function omitting the penalty term λK was utilized in Banerjee et al. (2005) in the context of finite mixture models.

Although this algorithm is straightforward, it has various drawbacks in practice. The most troublesome is that the functional dependency between the concentration parameter and the covariances destroys the rich-get-richer property of the DPMM because the counts of assignments to components $N_{k,-i}$ no longer influence which component gets assigned to an observed data point. Only the geometry in the data space matters. A new cluster is created by comparing the parameter λ against the distances between cluster centers and data points so that the number of clusters is controlled by the geometry alone, and not by the number of data points already assigned to each cluster. So, for high-dimensional datasets, it is not clear how to choose the parameter λ. By contrast, in the DPMM Gibbs sampler, the concentration parameter N_0 controls the rate at which new clusters are produced in a way which is, for fixed geometries, independent of the geometry. Another problem is that shrinking diagonal covariances to zero variance means that the component likelihoods become degenerate Dirac point masses which causes likelihood comparisons to be meaningless since the likelihood becomes infinite. So, we cannot choose parameters such as λ using standard model selection methods such as cross-validation.

While Jiang et al. (2012) can be seen as a nonparametric extension of the well known derivation of K-means from the (E-M) algorithm, Roychowdhury et al. (2013) have suggested an alternative SVA approach in the context of the iHMM. Herein we review their approach in the context of DPMMs and discuss their original formulation in Section 6.3. Instead of simply reducing the diagonal likelihood covariance to 0 variance in each dimension, Roychowdhury et al. (2013) represent the categorical distribution over the latent variables $z_1, \ldots, z_N$ in the more general exponential family form. The conditional distribution of the cluster indicator for point i given the mixture weights is given by:

$$p(z_i \,|\boldsymbol{\pi}) = \exp(-D_\phi(z_i, \boldsymbol{\pi}))\, b_\phi(z_i) \tag{4.4}$$

where $\boldsymbol{\pi} = (\pi_k)_{k=1}^{K}$ is the vector of mixture weights. Written in this form now we can also scale the variance of the categorical distribution over z. Furthermore, Roychowdhury et al. (2013) assume an additional dependency (which is not part of the DPMM) between the distribution of the cluster indicators and the component mixture distribution, in order for their diagonal variances to approach 0 simultaneously. That is, while Jiang et al. (2012) change the underlying DPMM structure only by assuming shrinking covariance, Roychowdhury et al. (2013) modify the underlying DPMM such that the conditional independence of the cluster parameters and cluster indicators no longer holds. Let us replace the distribution from Equation (4.4) with a scaled one:

$$p\left(z_i \left| \boldsymbol{\pi} \right.\right) = \exp\left(-\hat{\xi} D_\phi\left(z_i, \boldsymbol{\pi}\right)\right) b_{\tilde{\phi}}\left(z_i\right) \tag{4.5}$$

where $\tilde{\phi} = \hat{\xi}\phi$ which will keep the same mean as in Equation (4.4). Then following Roychowdhury et al. (2013) we assume that the likelihood $\tilde{F}\left(\tilde{\theta}\right)$ is scaled with ξ for which the equality $\hat{\xi} = \lambda_1 \xi$ holds for some real λ_1. Now, taking $\xi \to \infty$ would result in the appropriate scaling. After taking the limit and removing the constant terms we obtain the objective function of this new SVA approach:

$$\sum_{k=1}^{K} \sum_{i:z_i=k} D_\phi\left(\mathbf{x}_i, \boldsymbol{\mu}_k\right) + \lambda_1 D_\phi\left(z_i, \pi_k\right) + \lambda K \tag{4.6}$$

which is optimized with respect to (z, μ, π), and where $D_\phi\left(z_i, \pi_k\right) \propto -\log \pi_k$. Optimization with respect to the mixture weights results in the empirical probability for the cluster weights $\pi_k = \frac{N_k}{N}$. So, this objective function then can be rewritten as:

$$\sum_{k=1}^{K} \sum_{i:z_i=k} D_\phi\left(\mathbf{x}_i, \boldsymbol{\mu}_k\right) - \lambda_1 \log \frac{N_k}{N} + \lambda K \tag{4.7}$$

The E-M procedure that tries to optimize this objective function computes, for each observation $\mathbf{x}_i$, the K divergences to each of the existing clusters: $D_\phi\left(\mathbf{x}_i, \boldsymbol{\mu}_k\right)$ for $k = 1, \ldots, K$. Then, it takes into account the number of data points in each component by adjusting the corresponding divergence for cluster k by subtracting $\lambda_1 \log \frac{N_k}{N}$. After computing these adjusted distances, observation $\mathbf{x}_i$ is assigned to the closest cluster unless λ is smaller than all of these adjusted distances, in which case a new cluster is created. Then the cluster means are updated with the sample mean of observations assigned to each cluster, and in addition we now have to update the counts $N_1, \ldots, N_K$.

By contrast to the SVA algorithm proposed by Jiang et al. (2012), the SVA algoritm of Roychowdhury et al. (2013) no longer clusters the data purely on geometric considerations, but also takes into account the number of data points in each cluster. In this respect the method has greater flexibility, but at the same time, unlike MAP-DP, we can see that SVA algorithms do not actually-optimize the complete data likelihood of the original underlying probabilistic

DPMM which motivates their derivation. By assuming additional ad-hoc dependencies between the likelihood distribution and the distribution over the indicator variables, SVA algorithms effectively start from a different underlying probabilistic model which is not explicitly given. This makes them less principled and more heuristic than the MAP-DP algorithm we present here. So, while SVA algorithms are quite simple, they sacrifice several key statistical principals including structural interpretability and the existence of an underlying probabilistic generative model.

5. DPMM experiments

This section provides some empirical results so that we can compare the performance of MAP-DP against existing approaches.

5.1. Synthetic CRP parameter estimation

We examine the performance of the MAP-DP, collapsed Gibbs, DP-means (Kulis and Jordan, 2012) and variational DP (Blei and Jordan, 2004) on CRP-partitioned, non-spherical Gaussian data in terms of estimation error and computational effort. We generate 100 samples from a two-dimensional DPMM. The partitions are sampled from a CRP with fixed concentration parameter $N_0 = 3$ and data size $N = 600$. Gaussian component parameters are sampled from a *normal-Wishart* (NW) prior with parameters $\mu_0 = [2, 3]$, $c_0 = 0.5$, $\nu_0 = 30$, $\Lambda_0 = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$. This prior ensures a combination of both well-separated and overlapping clusters. We fit the model using MAP-DP, variational DP and Gibbs algorithms using the ground truth model values for the NW prior and the N_0 used to generate the data. Convergence for the Gibbs algorithm is tested using the Raftery diagnostic ($q = 0.025, r = 0.1, s = 0.95$) (Raftery and Lewis, 1992). We use a high convergence acceptance tolerance of $r = 0.1$ to obtain less conservative estimates for the number of iterations required. We use the most likely value from the Gibbs chain after burn-in samples (1/3 of the samples) have been removed.

Clustering estimation accuracy is measured using the *normalized mutual information* (NMI) metric (Vinh et al., 2010). The parameter λ for DP-means is set using a binary search procedure such that the algorithm gives rise to the correct number of partitions (see Appendix D). This approach favours DP-means as it is given knowledge of the true number of clusters which is not available to the other algorithms. For variational DP we set the truncation limit to ten times the number of clusters in the current CRP sample.

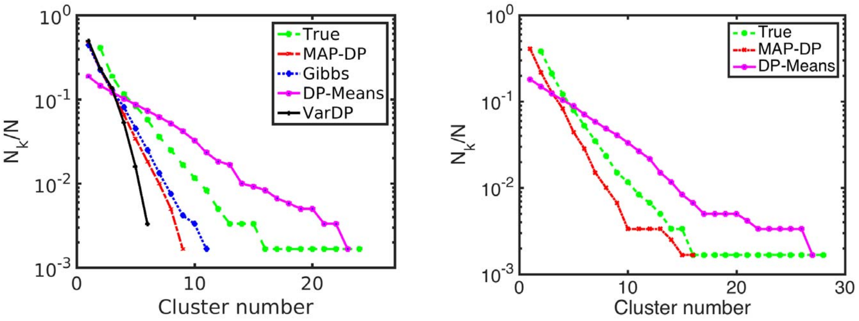
Both MAP-DP and Gibbs achieve similar clustering performance in terms of NMI whilst variational DP and DP-means have lower scores (Table 1). MAP-DP

Table 1
Performance of clustering algorithms on the CRP mixture experiment (Section 5.1). Mean and standard deviation (in brackets) reported across 100 CRP mixture samples. The range of the NMI is $[0, 1]$ *with higher values reflecting higher clustering accuracy.*

	Gibbs-MAP	MAP-DP	DP-means	Variational DP
Training set NMI	0.81 (0.1)	0.82 (0.1)	0.68 (0.1)	0.75 (0.1)
Iterations	1395 (651)	10 (3)	18 (7)	45 (18)

requires the smallest number of iterations to converge with the Gibbs sampler requiring, on average, 140 more iterations and DP-means 1.8 times. In Figure 5.1(a) the median partitioning[1] is shown in terms of the partitioning N_k/N and the number of clusters. As expected, when using a CRP prior, the sizes of the different clusters vary significantly with many small clusters containing only a few observations. MAP-DP and variational DP fail to identify the smaller clusters whereas the Gibbs sampler is able to do so to a greater extent. This is a form of underfitting where the algorithm captures the mode of the partitioning distribution but fails to put enough mass on the tails (the smaller clusters). The NMI scores do not reflect this effect as the impact of the smaller clusters on the overall measure is minimal. The poorer performance of the DP-means algorithm can be attributed to the non-spherical nature of the data as well as the lack of reinforcement effect that leads to underestimation of the larger clusters and overestimation of the smaller clusters.

This poor performance of DP-means is confirmed by modifying the CRP experiment to sample from spherical clusters (Figure 5.1(b)). The CRP is again sampled 100 times and the MAP-DP algorithm attains NMI scores of 0.88 (0.1) and DP-means scores NMI 0.73 (0.1). As the clusters are spherical, the lower performance of the DP-means algorithms is solely explained by the lack of reinforcement effect.



(a) Non-spherical (b) Spherical

Fig 5.1. *CRP mixture experiment; distribution of cluster sizes, actual and estimated using different methods. Cluster number ordered by decreasing size (horizontal axis) vs* $\frac{N_k}{N}$ *(vertical axis).*

[1] For each inference method, this is the median paritioning in terms of NMI out of the 100 DPMM sampled datasets.

Table 2

Clustering performance of DP-means, MAP-DP, and Gibbs samplers on UCI datasets, measured using NMI (two standard deviations in brackets), averaged over all runs. Higher NMI is better.

	DP-means	Gibbs	MAP-DP
Wine(178 observations, 13 dimensions)	0.42	0.71 (0.06)	0.86
Iris(150 observations, 4 dimensions)	0.76	0.75 (0.06)	0.76
Breast cancer(683 observations, 9 dimensions)	0.75	0.72 (0.01)	0.71
Soybean(266 observations, 35 dimensions)	0.36	0.45 (0.00)	0.40
Pima(768 observations, 8 dimensions)	0.03	0.14 (0.01)	0.07
Vehicle(846 observations, 18 dimensions)	0.21	0.10 (0.02)	0.15

5.2. UCI datasets

Next, we compare DP-means, MAP-DP and Gibbs sampling on six UCI machine learning repository datasets (Blake and Merz, 1998): *Wine*; *Iris*; *Breast cancer*; *Soybean*; *Pima* and *Vechicle*. We assess the performance of the methods using the same NMI measure as in Section 5.1. Class labels in the datasets are treated as cluster numbers.[2] There is either no or a negligibly small number of missing values in each of the data sets. The data types vary between datasets and features: Wine consists of integer and real data; Iris contains real data; Breast cancer consists of integer and categorical data; Soybean is categorical data; Pima is real data and Vehicle consists of integer data.

As in Section 5.1 we stop the Gibbs sampler using the Raftery diagnostic (Raftery and Lewis, 1992). For DP-means, we choose λ to give the true number of clusters in the corresponding dataset (Kulis and Jordan, 2012). For the Gibbs algorithm, we report the NMI of the most likely clustering from the whole chain of samples (Table 2). We also report the two standard deviations of the NMI computed at each sample of the chain after burn-in.

On almost all of the datasets (5 out of 6), MAP-DP is comparable to, or even better than, the Gibbs sampler, and on 4 out of 6 datasets it performs as well as or better than DP-means (Table 2). DP-means performs well on lower-dimensional datasets with a small number of clusters. In higher dimensions, it is more likely for the clusters to be elliptical rather than spherical and in such cases the other algorithms outperform DP-means because of the more flexible model assumptions. In addition, for higher dimensional data it is more often the case that the different features have different numerical scaling, so the squared Euclidean distance used in DP-means is inappropriate. Furthermore, MAP-DP and the Gibbs sampler are more robust to smaller clusters due to the longer tails of the *Student-T* predictive distribution (Appendix B) and the rich-get-richer effect of existing clusters assigned many observations. DP-means is particularly sensitive to geometric outliers and can easily produce excessive numbers of spurious clusters for poor choices of λ.

[2] We do not assess “Car” and “Balance scale” datasets used in Kulis and Jordan (2012) because they consist of a complete enumeration of 6 and 4 categorical factors respectively, and it is not meaningful to apply an unsupervised clustering algorithm to such a setting.

Table 3

Iterations required to achieve convergence for the DP-means and MAP-DP algorithm, and the Gibbs sampler, on datasets from the UCI repository.

	DP-means	Gibbs	MAP-DP
Wine	19	2,365	11
Iris	8	1,543	5
Breast cancer	8	939	8
Soybean	14	1059	9
Pima	20	1,189	17
Vehicle	12	939	9

Even though MAP-DP only gives a point estimate of the full joint posterior distribution, MAP-DP can in practice achieve higher NMI scores than for Gibbs due to MCMC convergence issues.

We emphasize that these algorithms attempt to maximize the model fit rather than maximize NMI. The true labels would not be available in practice and it is not always the case that maximizing the likelihood also maximizes NMI. Furthermore, if we choose the model hyperparameters for each dataset separately, by minimizing the negative log likelihood with respect to each parameter, higher NMI can been achieved, but choosing empirical estimates for the model parameters simplifies the computations.

In all cases, the MAP-DP algorithm converges more rapidly than the other algorithms (Table 3). The Gibbs sampler takes, typically, greater than 1000 more iterations than MAP-DP to achieve comparable NMI scores. The computational complexity per iteration for Gibbs and MAP-DP is comparable, requiring the computation of the same quantities. This makes the Gibbs sampler significantly less efficient than MAP-DP in finding a good labeling for the data. The computational price per iteration for DP-means can often be considerably smaller than MAP-DP or the Gibbs sampler, as one iteration often does not include a scan through all N data points. This occurs because the scan ends when a new cluster has to be created, unlike MAP-DP and Gibbs. But, this also implies that DP-means requires more iterations to converge than MAP-DP.

6. MAP-DP for infinite hidden Markov models

The simplicity of MAP-DP makes it straightforward to extend to more complex nonparametric models, such as the popular time series hidden Markov model. Mirroring the approach taken in Section 3, we can obtain an approximate MAP algorithm for the infinite HMM (Beal et al., 2002) (also known as the HDP-HMM (Teh et al., 2006)) for modeling sequential, time-series data.

HMMs can be seen as a generalization of finite mixture models where the cluster indicators that denote mixture component assignments are not independent of each other, but related through a Markov process. That is, each data point $\mathbf{x}_t$ of a sequence observations $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ is drawn independently of the other observations when conditioned on the state variable for time t, z_t. The state variables are linked through a state transition matrix where each row defines a mixture model for one of the values of the categorical distribution on the states. The current state z_t indexes a specific row of the transition matrix, with probabilities in this row serving as the mixing proportions for the choice of the next state z_{t+1}. Therefore the HMM does not involve a single mixture model, but rather a set of different mixture models, one for each value of the current state. (Beal et al., 2002) showed that if we replace the mixture models with a set of DPs, one for each value of the current state, a nonparametric variant of HMM is obtained that allows an unbounded set of states. In order to obtain sharing of available states across the sequence, the atoms associated with the state-conditional DPs are shared and so the transition matrix is modeled with an HDP (Teh et al., 2006).

Let us denote the base measure of the HDP by H where we restrict H to an exponential family distribution (for Bayesian conjugacy); N_0 and M_0 are *local* and *global* concentration parameters; z_{t-1} indicates the state chosen at time $t-1$. The HDP-HMM can then be written as:

$$\begin{aligned} G_0 &\sim \text{DP}\left(M_0, H\right) \\ G_{z_{t-1}} &\sim \text{DP}\left(N_0, G_0\right) \\ \vartheta_{tz_{t-1}} &\sim G_{z_{t-1}} \\ \mathbf{x}_t &\sim F\left(\vartheta_{tz_{t-1}}\right) \end{aligned}$$

where distribution over the data $F\left(\vartheta_{tz_{t-1}}\right)$ is a mixture distribution with mixing distribution $G_{z_{t-1}}$ over ϑ determined by the state pointed by z_{t-1}. We can also write the mixing distribution in terms of transition matrix $\boldsymbol{\pi}$ and base measure atoms θ, $G_{z_{t-1}} = \sum_{k=1}^{\infty} \pi_{z_{t-1}k}\delta_{\theta_k}$ and $\mathbf{x}_t \sim \sum_{k=1}^{\infty} \pi_{z_{t-1}} F\left(\theta_k\right)$ for all $G_1, \ldots, G_K$. The rows of the transition matrix $\boldsymbol{\pi}$ denote the mixture weights for each of the local DPs, while $\theta_1, \ldots, \theta_K$ are the shared atoms which are the same across $G_1, \ldots, G_K$, where K denotes the number of represented components, which in a nonparametric setting is unknown. The *global* DP is then characterized with $G_0 = \sum_{k=1}^{\infty} \kappa_k \delta_{\theta_k}$ where the elements of κ are the mixture weights of this global DP.

6.1. Gibbs sampler

As in the case of DPMMs (Section 2), exact inference is not available and Gibbs sampling is a standard choice for inference. A popular approach is based upon the *direct assignment* sampling approach for the HDP (Teh et al., 2006), where the cluster parameters θ and the transition matrix $\boldsymbol{\pi}$ are integrated out. The resulting Gibbs sampler then iterates between sampling the state indicators z and the global mixture weights κ. The mixture weights can be sampled from the corresponding Dirichlet posterior:

$$\kappa_1, \ldots, \kappa_K, \kappa_{K+1} \sim \text{Dirichlet}\left(M_1, \ldots, M_K, M_0\right) \tag{6.1}$$

where M_k counts how many times the transition to state k has been drawn from the global DP. In this representation, we do not keep the assignment variables for the global DP and some extra effort is needed to compute the global counts; we only use the assignments z_t of a point to its corresponding state and those assignments are used to compute the local DP counts, $N_{pk} = |\{t : z_{t-1} = p, z_t = k\}|$, that count how many times a transition occurred from state p to state k. Let m_{pk} be the count of how many times transitions from state p to state k have been drawn as a new transition from the global DP. We then have $M_k = \sum_p m_{pk}$. It is straightforward to see that if $N_{pk} \neq 0$ then $m_{pk} \neq 0$, because the first time the transition occurs from state p to state k, it is sampled from the global DP. Furthermore, the count m_{pk} will be limited by N_{pk}, as at most all transitions from state p to state k are sampled from the global DP. We can then use the *Polya urn* sampling scheme underlying the transition probability from state p to state k to sample m_{pk}:

$$p\left(z_{t+1} = k \left| z_t = p\right.\right) \propto \begin{cases} N_{pk}^{-z_{t+1}} - 1 & \text{for an existing transition from } p \text{ to } k \\ N_0 \kappa_k & \text{for a new transition from } p \text{ to } k \end{cases} \tag{6.2}$$

Due to the exchangeability of rows in the transition matrix, we can sample from Equation (6.2) sequentially N_{pk} times, gradually increasing N_{pk}, and keeping count of how many times the transition from the second term has been sampled. The recorded count m_{pk} is unbiased and can be marginalized over p to obtain M_k (Van Gael, 2012).

In order to update the state indicators, for $t = 1, ..., T$ we sample z_t from the corresponding probability:

$$p\left(z_t \left| \mathbf{X}, z_{-t}, \kappa\right.\right) \propto p\left(\mathbf{x}_t | z_{-t}, z_t, \mathbf{X}\right) p\left(z_t \left| z_{-t}, \kappa\right.\right) \tag{6.3}$$

where z_{-t} denotes all indicators excluding the one for data point t. The first term is the posterior predictive distribution of data point $\mathbf{x}_t$ given its state and it will be obtained after integrating out the state parameters θ assuming a conjugate prior for the exponential family likelihood. The resulting distribution is computed in the same way as in Section 3.1:

$$\begin{aligned} p\left(\mathbf{x}_t | \dots\right) \propto \exp &\left(\psi_0\left(\tau + \sum_{j:z_j=k} g\left(\mathbf{x}_j\right), \eta + N_k\right)\right. \\ &\left. -\psi_0\left(\tau + \sum_{j:z_j=k, j\neq t} g\left(\mathbf{x}_j\right), \eta + N_{k,-t}\right)\right) \end{aligned} \tag{6.4}$$

To compute the second term recall that DPs for each row of the transition matrix are independent, given the shared base measure κ. Then if $N_{k\cdot}$ denotes number of transitions from state k and $N_{\cdot k}$ number of transitions to state k we can write:

$$p\left(z_t = k\,|z_{-t}, \kappa\right) \propto \begin{cases} \left(N^{-t}_{z_{t-1}k} + N_0\kappa_k\right) \frac{\left(N^{-t}_{kz_{t+1}} + N_0\kappa_{z_t+1}\right)}{N^{-t}_{k\cdot} + N_0} & \text{for } k \leq K,\ z_{t-1} \neq k \\ \left(N^{-t}_{z_{t-1}k} + N_0\kappa\right) \frac{\left(N^{-t}_{kz_{t+1}} + 1 + N_0\kappa_{z_t+1}\right)}{N^{-t}_{k\cdot} + N_0 + 1} & \text{for } z_{t-1} = z_{t+1} = k \\ \left(N^{-t}_{z_{t-1}k} + N_0\kappa_k\right) \frac{\left(N^{-t}_{kz_{t+1}} + N_0\kappa_{z_t+1}\right)}{N^{-t}_{k\cdot} + N_0 + 1} & \text{for } z_{t-1} = k \neq z_{t+1} \\ N_0\kappa_k\kappa_{z_t+1} & \text{for } k = K + 1 \end{cases} \tag{6.5}$$

6.2. MAP-DP for iHMMs

In our proposed iterative MAP approach, we sweep through the latent variables and at each iteration update them one at a time with their respective MAP values. Following the direct assignment construction presented in the previous section, the random variables to be updated are the global DP mixture weights κ and the state indicators z. The mode of the Dirichlet posterior is available in closed form for concentration parameter $M_0 \geq 1$, so we can update κ using:

$$\begin{array}{rlc} \kappa_k = \frac{M_k - 1}{M_\cdot - K - 1} & \text{for an existing state } k = 1, \dots, K \\ \kappa_{K+1} = \frac{M_0 - 1}{M_\cdot - K - 1} & \text{for a new state} \end{array} \tag{6.6}$$

where $M_\cdot = \sum_{k=0}^{K} M_k$. The counts m_{pk} can be computed by numerical optimization of the corresponding distribution over partitions provided in (Antoniak, 1974). However, the expression involves Stirling numbers of the first kind which may be numerically challenging to compute when the method is applied to time series with a large number of observations. One approach to avoid this issue is using Equation (6.2) to sequentially compute the corresponding counts.

In order to update the state indicators, we sweep through each observation $\mathbf{x}_t$ and then compute the negative log probability for each existing state k and for a new state $K + 1$:

$$\begin{aligned} q_{t,k} &= - \log p\left(\mathbf{x}_t\,|z_t = k, \mathbf{X}_{-t}\right) - \log p\left(z_t = k\,|z_{-t}, \kappa\right) \\ q_{t,K+1} &= \qquad - \log p\left(\mathbf{x}_t\,|\tau, \eta\right) - \log\left(N_0\kappa_{K+1}\kappa_{z_t+1}\right) \end{aligned}$$

where again and without losing generality, we can omit the terms independent of k. For each observation in the time series, we compute the $K + 1$-dimensional vector q_t and select the state number according to:

$$z_t = \underset{k \in \{1,\dots,K,K+1\}}{\arg\min}\, q_{t,k} \tag{6.7}$$

As with the MAP-DP case, the algorithm proceeds to the next point in the time series $\mathbf{x}_{t+1}$ by updating the state sufficient statistics to reflect the new value of the cluster assignment z_t and remove the effect of data point $\mathbf{x}_{t+1}$. The scheme is still guaranteed to converge to a fixed point solution as each step does not increase the NLL. However, when the Polya urn scheme in Equation (6.2) is used to compute the global DP counts its stochastic nature will result in minor fluctuations in the model NLL. The algorithm still falls into a local optima, but minor fluctuations occur between iterations. As the NLL is significantly more influenced by the likelihood terms $\log p\left(\mathbf{x}_t \left| z_t = k, \mathbf{X}_{-t}\right.\right)$ and the assignment probabilities, stopping the scheme at the local minima is straightforward.

6.3. SVA for iHMMs

Jiang et al. (2012) propose an alternative inference algorithm for efficient estimation of iHMMs based on the SVA approach (Section 4). Mirroring the simplified inference assumptions explained in Section 4, Jiang et al. (2012) have extended the SVA approach to HDP mixtures and so can be readily applied for inference in the HDP-HMM model. The resulting algorithm extends the objective function derived for the simpler DPMM with an additional term penalizing the number of transitions leading out of each state:

$$\min_{z,\mu,K} \sum_{t=1}^{T} \sum_{k=1}^{K^{\text{global}}} D_{\phi}\left(\mathbf{x}_t, \mu_k\right) + \lambda_1 \sum_{p=1}^{K^{\text{global}}} K_p^{\text{local}} + \lambda_2 K^{\text{global}} \tag{6.8}$$

In the case of the iHMM, K_p^{local} denotes the number of states for which a transition from state p exists, and K^{global} denotes the total number of represented states in the finite time series. Then λ_2 penalizes the creation of new states, while λ_1 controls how likely new transitions are between existing states. As in the DPMM SVA algorithm (Section 4), standard model selection techniques cannot be applied to select the values of λ_1 and λ_2 and reinforcement terms have been stripped away. Alternatively, introducing some additional assumptions which imply an entirely ad-hoc coupling between the state indicator distribution and the data likelihood probabilities, (Roychowdhury et al., 2013) derives an SVA algorithm that optimizes the following objective function:

$$\min_{K,Z,\mu,T} \sum_{t=1}^{T} \sum_{k=1}^{K^{\text{global}}} D_{\phi}\left(\mathbf{x}_t, \mu_k\right) - \lambda \sum_{t=2}^{T} \sum_{k=1}^{K^{\text{global}}} \log \frac{N_{z_{t-1}k}}{N_{z_{t-1}\cdot}} + \lambda_1 \sum_{p=1}^{K^{\text{global}}} K_p^{\text{local}} + \lambda_2 K^{\text{global}} \tag{6.9}$$

(Roychowdhury et al., 2013) also introduces a further algorithm, unrelated to Gibbs sampling, that attempts to optimize this objective function whilst also taking advantage of dynamic programming. The method is referred to as asymp-iHMM in our synthetic experiment below.

6.4. Synthetic study

We compare the performance of the Gibbs, SVA and MAP-DP algorithms for iHMMs on synthetically generated 3-dimensional data from a five-state HMM with Gaussian data models. The five Gaussian components are parameterized

Table 4
NMI and number of iterations to convergence for different iHMM inference algorithms. '+' indicates convergence was not obtained.

	NMI	Iterations
Gibbs sampler	0.77	2500+
asymp-iHMM	0.58	12
MAP-iHMM	0.62	13

with isotropic means $\left[\mu_1^d, \mu_2^d, \mu_3^d, \mu_4^d, \mu_5^d\right] = [0.2, 1.1, 1.8, 3.4, 4.8]$ for $d = 1, 2, 3$, each with shared spherical covariance $\sigma\mathbf{I}_3$ where $\sigma = 0.9$. At each time step the HMM has 0.96 probability of self-transition and equal probability to transition to any of the remaining four states. 4000 data points are generated and performance of the algorithms is measured using the NMI between the true and estimated state assignments. The Gibbs sampler is evaluated for the state assignments that maximize the model likelihood. As shown in Table 4, MAP-iHMM performs similarly to asymp-iHMM from (Roychowdhury et al., 2013) whilst keeping the underlying probabilistic iHMM model intact and retaining a non-degenerate complete data likelihood.

Inferring states in this data is difficult due to overlapping observation likelihood models, and we find that the Gibbs sampler significantly outperforms both point estimation approaches, but at an even greater computational cost than for the simpler DPMM. This example illustrates the computational challenges of using MCMC inference for more complex hierarchical models. In Figure 6.1 we observe that the state sequence obtained by MAP and SVA is similar and underestimates the true number of states.

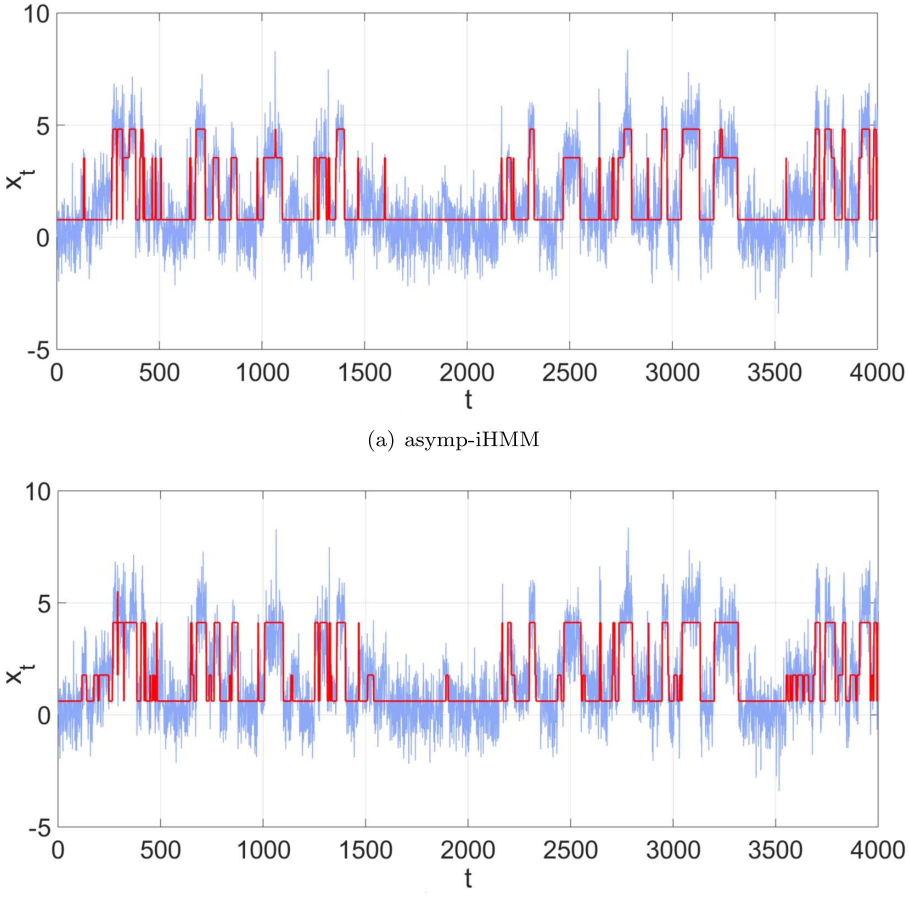
7. MAP-DP for semiparametric mixed effects models

Hierarchical modeling is commonly used in the analysis of longitudinal health data. A particular model that is widely used in practice is the *linear mixed effects model*:

$$\begin{aligned}\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta}_i + \epsilon_i \\ \boldsymbol{\beta}_i &\sim P\end{aligned} \tag{7.1}$$

where $\mathbf{y}_i$ the observation vector for individual $i \in \{1, \ldots, N\}$, $\epsilon_i \sim \mathcal{N}\left(0, \tau_\sigma^{-1}\mathbf{I}\right)$ is the subject-specific observation noise with τ_σ the within-subject precision and P the distribution of the *random effects* $\boldsymbol{\beta}_i$ (Dunson, 2010). $\mathbf{X}_i$ are the inputs for the random effects $\boldsymbol{\beta}_i$ and the fixed effect regression parameters are equal to the mean of the distribution P. The distribution P is commonly specified to be Gaussian due to analytical tractability and computational simplicity. However, the assumption of normality is seldom justified and the assumptions of symmetry and unimodality are often found to be inappropriate (Dunson, 2010).

Semiparametric mixed effects models have been proposed to relax the normality assumption by placing a DPMM prior on P (Kleinman and Ibrahim, 1998). However, inference for such models is usually performed using MCMC



(a) asymp-iHMM

(b) MAP-iHMM

Fig 6.1. *Synthetically generated HMM data is in blue. The red line is the estimated state centroids associated with each point, where centroids have been estimated using (a) asymp-iHMM and (b) MAP-iHMM.*

requiring large computational resources and careful tuning of algorithmic parameters. This makes MCMC approaches particularly difficult to implement on large data sets. The increasing availability of large longitudinal data sets warrants the investigation of computationally efficient inference approaches such as MAP-DP. Here in order to construct a semiparametric mixed effects model, we will use an iterative MAP algorithm similar to MAP-DP above with the only difference of not integrating out the component parameters.

We construct the model by first placing a DPMM prior on $\boldsymbol{\beta}_i$ in Equation (7.1). As we are interested in the interpretation of the clusters we do not collapse out the cluster parameters and the update steps described for MAP-DP are slightly altered (as in non-collapsed Gibbs) where the random effects $\boldsymbol{\beta}_i$ are substituted for the individual data points $\mathbf{x}_i$; an additional step updating the component means and precisions also needs to be included. Two further steps are needed to update the random effects $\boldsymbol{\beta}_i$ and within-subject precision τ_σ. The conditional $p\left(\boldsymbol{\beta}_i | \tau_\sigma, z_i = k, \boldsymbol{\mu}_k, \mathbf{R}_k\right)$ for the random effects $\boldsymbol{\beta}_i$ is:

$$\mathcal{N}\left(\boldsymbol{\beta}_i \left| \left(\tau_\sigma \mathbf{X}_i^T \mathbf{X}_i + \mathbf{R}_k\right)^{-1} \left(\tau_\sigma \mathbf{X}_i^T \mathbf{y}_i + \mathbf{R}_k \boldsymbol{\mu}_k\right), \left(\tau_\sigma \mathbf{X}_i^T \mathbf{X}_i + \mathbf{R}_k\right)^{-1}\right.\right) \tag{7.2}$$

where the conditioning is on the assigned cluster k with mean $\boldsymbol{\mu}_k$ and precision $\mathbf{R}_k$. We place a conjugate Gamma prior on the within-subject precision $\tau_\sigma \sim \text{Gamma}\left(a_{\sigma^2}, b_{\sigma^2}\right)$ allowing for the calculation of the conditional posterior:

$$\begin{aligned} &p\left(\tau_\sigma | \boldsymbol{B}, a_{\sigma^2}, b_{\sigma^2}\right) \\ &\quad = \text{Gamma}\left(\tau_\sigma \left| a_{\sigma^2} + \frac{N}{2}, b_{\sigma^2} + \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_i\right)^T \left(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}_i\right)\right.\right) \end{aligned} \tag{7.3}$$

where $\boldsymbol{B}$ is the collection of all random effects $\left(\boldsymbol{\beta}_i\right)_{i=1}^{N}$. The modes of both conditionals needed for MAP-DP are easily calculated in addition to the negative log likelihood necessary to check for convergence.

7.1. English longitudinal survey of ageing

We apply the semiparametric mixed effects model above to the *English longitudinal survey of ageing* (ELSA), a large longitudinal survey of older adults aged over 50 in the United Kingdom. ELSA is a multi-purpose health study which follows individuals aged 50 years or older (Netuveli et al., 2006). Collected health-related factors include clinical, physical, financial and general well-being. Of primary interest is the effect of the different factors on *quality of life* (QoL) measured using a compound measure of several health and socio-economic indicators. The ELSA survey has been conducted in five waves spanning ten years. In this preliminary study we look at the response of 6,805 individuals across all 5 waves.

We wish to check the hypothesis that measures of cognition such as memory and executive mental function, as estimated by verbal fluency, are useful predictors of QoL and whether they are more informative than standard measures of depression and *activities of daily living* (ADL) that have been found to be statistically significant predictors of QoL (Netuveli et al., 2006). We propose to answer these two questions via selection of two models with different sets of covariates. The first model includes depression and ADL as inputs whereas the second model includes measures of cognitive ability, specifically prospective memory and verbal fluency. The models are assessed using 5-fold cross-validation and computing the average held-out likelihood (Equation (3.4) in Section 3.2).

The model that includes ADL and depression as covariates achieves a significantly lower average held-out likelihood than the competing model containing cognitive measures suggesting that ADL and depression are more informative predictors of QoL than the cognitive measures we considered (Table 5).

Table 5
Cross-validation average held-out likelihood for two models.

Cognitive measures	Depression + ADL
0.364	3.834

The average elapsed time for fitting the models using MAP inference is 11.05 seconds and 16.29 seconds[3]. For comparison we performed inference using a truncated DP random effects model with MCMC and 100,000 iterations to ensure convergence on less than half of the data (3,000 individuals) and the resulting time to convergence is in excess of five hours making inference on larger data sets impractical. On the other hand, the rapid inference obtained using MAP-DP enables a wide array of diagnostic and validation methods to be exploited, which suggests the approach can be scaled up to very large datasets.

8. Discussion and future directions

We have presented a simple algorithm for inference in DPMMs based on non-degenerate MAP, and demonstrated its efficiency and accuracy by comparison to the ubiquitous Gibbs sampler, variational DP, and the SVA algorithms. The attractiveness of SVA lies in the simplicity and scalability of the resulting algorithms but as we have shown, it entails significant structural departures from the DPMM as well as removing from the modeler's arsenal standard tools of model comparison and selection. We believe our approach is highly relevant to applications since, unlike SVA, it retains the preferential attachment (rich-get-richer) property while needing two orders of magnitude fewer iterations than Gibbs. Unlike SVA, the out-of-sample likelihood may be computed allowing the use of standard model selection and model fit diagnostic procedures. Lastly, this non-degenerate MAP approach does not require the approximation inherent to the factorization assumptions of VB.

As with most MAP methods, MAP-DP can get trapped in local minima, however, standard heuristics such as multiple random restarts can be employed to mitigate this risk. This would increase the total computational cost of the algorithm somewhat but even with random restarts it would still be far more efficient than the Gibbs sampler.

Although not reported here due to space limitations, we point out that different implementations of the Gibbs sampler can lead to different MAP inference algorithms for DPMMs. For example, different MAP procedures can be derived from the different Gibbs samplers presented in (Neal, 2000b). In general, we have found these alternative algorithms to be less robust in practice, as they do not integrate over the uncertainty in the cluster parameters. However, when such assumptions are justified, our MAP approach can be readily applied to different constructions of the DPMM, for example to allow for non-conjugate choice of priors (extending Algorithm 7 (Neal, 2000b)).

[3] The reported runtimes for MAP-DP and MCMC were obtained on Matlab R2013a (8.1.0.604) 64-bit (glnxa64), i7-2600 CPU with 3.40GHz processor, ubuntu PC.

The simplicity of MAP-DP allows us to easily extend the algorithm to more complex, composite models such as the iHMM as demonstrated in Section 6 or semiparametric hierarchical mixed-effects models in Section 7. On sequential time series data, MAP-DP does as well as the hybrid SVA approach of Roychowdhury et al. (2013) in terms of accuracy and computational load whilst avoiding many of its limitations. We have also contrasted the MAP and SVA approaches from a theoretical perspective and highlighted some of the inherent theoretical and practical limitations of the latter.

The generality and the simplicity of MAP-DP makes it reasonable to adapt to other Bayesian nonparametric mixture models, for example the *Pitman-Yor process* which generalizes the CRP (Pitman and Yor, 1997). The MAP approach can also be applied to hierarchical BNP models such as the nested DP (Rodriguez et al., 2008). Another useful direction, for large-scale datasets in particular, would be to extend our approach to perform inference that does not need to sweep through the entire dataset in each iteration, for increased efficiency (Welling and Teh, 2011).

Appendix A: Estimating the model hyperparameters

In Bayesian models, we would ideally like to choose our hyperparameters (θ_0, N_0) where $\theta_0 = (\boldsymbol{\tau}, \eta)$ using some additional information that we have for the data. This could be related to the way data is collected, the nature of the data itself, or expert knowledge about the problem at hand. For instance, when there is prior knowledge on the number of clusters, the concentration parameter N_0 could be set using the fact that the prior expected number of clusters for a DP is $N_0 \log N$.

In cases where this is not feasible, we have considered the following alternatives:

1. *Empirical Bayes*. Set the hyperparameters to their corresponding maximum marginal likelihood values. The maximum marginal likelihood expression for θ_0 will be different for different data types and will not always be available in closed form.
2. *Multiple restarts*. Run MAP-DP with different starting values for each of the hyper parameters (θ_0, N_0), compute the negative log likelihood, and change one of the hyperparameters while holding the rest fixed. Then, restart MAP-DP with the prior parameter. Set that hyperparameter to the value resulting in the smallest negative log likelihood and proceed in the same way for the next hyperparameter of the model. *Bayesian optimisation* (BO) (Snoek et al., 2012) has also been proposed to fit model hyperparameters but requires the specification of a Gaussian process and associated priors for this that may be challenging in practice. We have therefore not utilised this approach and prefer the simpler greedy search. However in certain cases BO may be more efficient in terms of the number of MAP-DP iterations required.

3. *MAP estimate.* Place a prior on the hyperparameter and numerically compute the mode of that posterior. For instance, using a gamma prior on N_0, $p\left(N_0\right) = \text{Gamma}\left(a_{N_0}, b_{N_0}\right)$, the posterior is proportional to

$$p\left(N_0|N,K\right) \propto \frac{\Gamma\left(N_0\right)}{\Gamma\left(N_0+N\right)} N_0^{K+a_{N_0}-1} \exp\left(-b_{N_0}N_0\right) \tag{A.1}$$

   We can numerically minimize the negative log of this posterior using Newton's method. To ensure the solution is positive we compute the gradient with respect to $\log N_0$: as Rasmussen (1999) notes $p\left(\log N_0|N,K\right)$ is log-concave and therefore has a unique maximum.
4. *Cross-validation.* By considering a finite set of values for $\left(\theta_0, N_0\right)$, choose the value corresponding to the maximum average out-of-sample likelihood across all cross-validation repetitions (see Section 3.2). This approach is taken in Blei and Jordan (2004) to compare different inference methods.

We have found the second approach above to be the most effective where empirical Bayes can be used to obtain the values of the hyperparameters at the first run of MAP-DP. For small datasets we recommend using the cross-validation approach as it can be less prone to overfitting.

Appendix B: Predictive distribution functions

In MAP-DP, the computation requires the collapsed prior predictive distribution $p\left(\mathbf{x}|\boldsymbol{\tau},\eta\right)$, and also the collapsed posterior predictive distribution $p\left(\mathbf{x}|\boldsymbol{\tau}_{k,-i},\eta_{k,-i}\right)$. These predictive distributions require the updated cluster posterior hyper parameters. These updates depend upon the distribution, and the data type, of each data point $\mathbf{x}_i$. When the distribution is from the *exponential family*, the prior distribution over the parameters can be chosen to be *conjugate*: the prior over the parameters of the data distribution and the posterior have the same form of distribution. This simplifies the hyper parameter updates, and, furthermore, the form of the prior and posterior predictive distributions is the same and is available in closed form. The table below lists some possible data types and distributions, their conjugate prior/posterior distribution, the names given to the hyper parameters and the corresponding name of the predictive distributions.

Distribution of data $\mathbf{x}_i$	Data type	Conjugate prior/posterior	Parameters	Predictive distribution
Spherical normal (known variance)	$\mathbf{x} \in \mathbb{R}^D$	Spherical normal	$\left(\boldsymbol{\mu}, \sigma^2\right)$	Spherical normal
Multivariate normal (known covariance)	$\mathbf{x} \in \mathbb{R}^D$	Multivariate normal	$\left(\boldsymbol{\mu}, \mathbf{\Sigma}\right)$	Multivariate normal
Multivariate normal	$\mathbf{x} \in \mathbb{R}^D$	Normal-Wishart (NW)	$\left(\boldsymbol{m}, c, \mathbf{B}, a\right)$	Multivariate Student-T
Exponential	$x \in \mathbb{R}$, $x \geq 0$	Gamma	$\left(\alpha, \beta\right)$	Lomax
Categorical	$x \in \{1, 2, \ldots D\}$	Dirichlet	$\left(\alpha_1, \ldots, \alpha_D\right)$	Dirichlet multinomial

Distribution of data $\mathbf{x}_i$	Data type	Conjugate prior/posterior	Parameters	Predictive distribution
Binomial	$x \in \{0, 1, \dots n\}$	Beta	(α, β)	Beta-binomial
Poisson	$x \in \mathbb{Z}$, $x \geq 0$	Gamma	(α, β)	Negative-binomial
Geometric	$x \in \mathbb{Z}$, $x \geq 0$	Beta	(α, β)	Ratio of beta functions

In the examples presented in this paper the data likelihood is multivariate Gaussian and we describe this case in more detail. Specifically when each data point $\mathbf{x} \in \mathbb{R}^D$ is assumed to be multivariate Gaussian with unknown mean vector and precision matrix, the conjugate prior distribution of the Gaussian parameters is NW, with hyperparameters $\theta_0 = (\boldsymbol{m}_0, c_0, \mathbf{B}_0, a_0)$. Then, the posterior distribution for each cluster is also NW, with hyperparameters $\theta_k^{-i} = \left(\boldsymbol{m}_k^{-i}, c_k^{-i}, \mathbf{B}_k^{-i}, a_k^{-i}\right)$. These are updated for each cluster according to:

$$\begin{aligned}
\boldsymbol{m}_k^{-i} &= \frac{c_0 \boldsymbol{m}_0 + N_{k,-i} \bar{\mathbf{x}}_{k,-i}}{c_0 + N_{k,-i}} \\
c_k^{-i} &= c_0 + N_{k,-i} \\
\mathbf{B}_k^{-i} &= \left(\mathbf{B}_0^{-1} + \mathbf{S}_{k,-i} + \frac{c_0 N_{k,-i}}{c_0 + N_{k,-i}} \left(\bar{\mathbf{x}}_{k,-i} - \boldsymbol{m}_0\right) \left(\bar{\mathbf{x}}_{k,-i} - \boldsymbol{m}_0\right)^T \right)^{-1} \\
a_k^{-i} &= a_0 + N_{k,-i}
\end{aligned} \tag{B.1}$$

where:

$$\begin{aligned}
\bar{\mathbf{x}}_{k,-i} &= \frac{1}{N_{k,-i}} \sum_{j: z_j = k, j \neq i} \mathbf{x}_j \\
\mathbf{S}_{k,-i} &= \sum_{j: z_j = k, j \neq i} \left(\mathbf{x}_j - \bar{\mathbf{x}}_{k,-i}\right) \left(\mathbf{x}_j - \bar{\mathbf{x}}_{k,-i}\right)^T
\end{aligned} \tag{B.2}$$

The predictive distributions $p\left(\mathbf{x} | \boldsymbol{\tau}, \eta\right)$ and $p\left(\mathbf{x} | \boldsymbol{\tau}_{k,-i}, \eta_{k,-i}\right)$ are D-dimensional multivariate Student-T distributions, whose negative log, written in terms of the parameters $(\mu, \mathbf{\Lambda}, \nu)$ is:

$$\begin{aligned}
-\log f\left(\mathbf{x}|.\right) &= \frac{\nu + D}{2} \log \left[1 + \nu^{-1} \left(\mathbf{x} - \boldsymbol{\mu}\right)^T \mathbf{\Lambda} \left(\mathbf{x} - \boldsymbol{\mu}\right) \right] \\
& \quad - \frac{1}{2} \log |\mathbf{\Lambda}| + \log \Gamma \left(\frac{\nu}{2} \right) + \frac{D}{2} \log \left(v \pi \right) - \log \Gamma \left(\frac{\nu + D}{2} \right)
\end{aligned}$$

where the Student-T parameters $(\boldsymbol{\mu}, \mathbf{\Lambda}, \nu)$ are given in terms of the NW parameters $\boldsymbol{\mu} = \boldsymbol{m}$, $\nu = a - D + 1$ and $\mathbf{\Lambda} = \frac{c\nu}{c+1} \mathbf{B}$. We note that fast incremental updates of all these parameters are possible when including and then removing a single data point from a cluster, see Raykov et al. (2014) for details.

Appendix C: Bregman divergences

The conditional probabilities for the DPMM can be expressed using the general *distortion* measure known as *Bregman divergence* (Banerjee et al., 2005). The

Bregman divergence between any two vectors $\mathbf{x}$ and θ is defined as $D_\phi(\mathbf{x},\theta) = \phi(\mathbf{x}) - \phi(\theta) - \langle \mathbf{x}-\theta, \nabla\phi(\theta)\rangle$ for the function $\phi : S \to \mathbb{R}$ being differentiable and strictly convex on a closed convex set $S \subseteq R^D$. Bregman divergences can be efficiently used to provide a compact parameterization of exponential family distributions with their expectation parameter. This generalizes the result that a group of points are summarized by their mean in Euclidean space to all spaces that can be described with Bregman divergence as a distortion measure.

Appendix D: DP-means λ parameter binary search

In our experiments with the DP-means algorithm, it is necessary to have an automatic way of obtaining the parameter λ for synthetic experiments where we wish to obtain a specific number of clusters K_{target}. We use a binary search approach where λ is set in a sequence of binary search steps:

1. *Initialisation*: Set λ to the mid-point of the range $\left[L_1 = 0, U_1 = M^2\right]$ where L_1, U_1 are respectively the lower and upper bounds of the range for the first iteration. M^2 is the maximal squared Euclidean distance and is set to $M^2 = \sum_{d=1}^{D}\left(\max(\boldsymbol{x}_d) - \min(\boldsymbol{x}_d)\right)^2$ where $\max(\boldsymbol{x}_d)$, $\min(\boldsymbol{x}_d)$ are respectively the upper and lower bounds of the data for dimension d. (The use of the maximal Euclidean distance originates in the DP-means algorithm step which creates a new cluster when $d_{i,k} > \lambda$ where $d_{i,k}$ is the squared Euclidean distance of data point i to the mean of cluster k.)
2. *For iteration* $i = 1, 2, \ldots$
   (a) Run the DP-means algorithm with $\lambda = \frac{1}{2}(U_i + L_i)$ which returns K_{obtained},
   (b) If $K_{\text{obtained}} > K_{\text{target}}$ then there are too many clusters so we will increase λ. We update the lower bound $L_{i+1} = \lambda$ and leave the upper bound unchanged $U_{i+1} = U_i$,
   (c) If $K_{\text{obtained}} < K_{\text{target}}$ there are too few clusters so we need to decrease λ. We update the upper bound $U_{i+1} = \lambda$ and leave the lower bound unchanged $L_{i+1} = L_i$,
   (d) Stop when $K_{\text{obtained}} = K_{\text{target}}$.

References

Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174, 1974. MR0365969

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, December 2005. ISSN 1532-4435. MR2249870

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. MR2091634

Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden Markov model. In *Machine Learning*, pages 29–245. MIT Press, 2002.

Christopher Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006. ISBN 0387310738. MR2247587

David Blackwell. Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18(1):105–110, 03 1947. URL http://dx.doi.org/10.1214/aoms/1177730497. MR0019903

Catherine Blake and Christopher J. Merz. {UCI} repository of machine learning databases. 1998.

David Blei and Michael I. Jordan. Variational methods for the Dirichlet process. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, page 12, 2004.

Olivier Bousquet and Léon Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1727–1735, 2013a.

Tamara Broderick, Brian Kulis, and Michael I. Jordan. Mad-bayes: Map-based asymptotic derivations from bayes. In *ICML (3)*, pages 226–234, 2013b.

Wang Chong, John W. Paisley, and David M. Blei. Online variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.

David B. Dahl. Modal clustering in a class of product partition models. *Bayesian Analysis*, 4(2):243–264, 2009. MR2507363

Hal Daumé. Fast search for Dirichlet process mixture models. In *International Conference on Artificial Intelligence and Statistics*, 2007.

David Dunson. Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics: Principles and Practice*. Cambridge University Press, 2010. MR2730665

Thomas Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 1(2):209–230, 03 1973. URL http://dx.doi.org/10.1214/aos/1176342360. MR0350949

Michael C. Hughes and Erik B. Sudderth. Memoized online variational inference for Dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 1133–1141, 2013.

Michael C. Hughes, Dae Il Kim, and Erik B. Sudderth. Reliable and scalable variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics*, pages 370–378, 2015.

Ke Jiang, Brian Kulis, and Michael I. Jordan. Small-variance asymptotics for exponential family Dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 3158–3166, 2012.

Josef Kittler and Janos Föglein. Contextual classification of multispectral pixel data. *Image and Vision Computing*, 2(1):13–29, 1984.

Ken Kleinman and Joseph Ibrahim. A semiparametric Bayesian approach to the random effects model. *Biometrics*, 54(3):921–938, 1998.

Brian Kulis and Michael I. Jordan. Revisiting K-means: New algorithms via Bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 513–520, 2012.

Jeffrey W. Miller and Matthew T. Harrison. A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pages 199–206, 2013.

Radford Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000a. MR1823804

Radford Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000b. MR1823804

Gopalakrishnan Netuveli, Richard Wiggins, Zoe Hildon, Scott Montgomery, and David Blane. Quality of life at older ages: Evidence from the english longitudinal study of aging (wave 1). *Journal of Epidemiology and Community Health*, 60(4):357–363, 2006.

Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995. MR1337249

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997. URL `http://dx.doi.org/10.1214/aop/1024404422`. MR1434129

Adrian Raftery and Steven Lewis. How many iterations in the Gibbs sampler? *Bayesian Statistics*, 4(2):763–773, 1992.

Carl Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*, pages 554–560, 1999.

Yordan Raykov, Alexis Boukouvalas, and Max A. Little. Simple approximate MAP inference for Dirichlet processes. *arXiv:1411.0939*, 2014.

Abel Rodriguez, David Dunson, and Alan Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483), 2008. MR2528831

Anirban Roychowdhury, Ke Jiang, and Brian Kulis. Small-variance asymptotics for hidden Markov models. In *Advances in Neural Information Processing Systems*, pages 2103–2111, 2013.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

Y. W. Teh, K. Kurihara, and M. Welling. Collapsed variational inference for the HDP. In *Advances in Neural Information Processing Systems*, pages 1481–1488, 2008.

Yee Teh, Michael I. Jordan, Matthew Beal, and David Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006. MR2279480

Jurgen Van Gael. *Bayesian Nonparametric Hidden Markov Models*. PhD thesis, University of Cambridge, 2012.

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 1088–1095, 2008.

Nguyen Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010. MR2738784

Lianming Wang and David Dunson. Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216, 2011. MR2816545

Max Welling and Kenichi Kurihara. Bayesian K-means as a maximization-expectation algorithm. In *SDM*, pages 474–478. SIAM, 2006. MR2337960

Max Welling and Yee Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 681–688, 2011.

Xiaole Zhang, David J. Nott, Christopher Yau, and Ajay Jasra. A sequential algorithm for fast fitting of Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 23(4):1143–1162, 2014. MR3270715