Electronic Journal of Statistics
Vol. 3 (2009) 507–545
ISSN: 1935-7524
DOI: 10.1214/08-EJS339

Mean field inference for the Dirichlet process mixture model

O. Zobay

Department of Mathematics, University of Bristol,
University Walk, Bristol BS8 1TW, United Kingdom

Abstract: We present a systematic study of several recently proposed methods of mean field inference for the Dirichlet process mixture (DPM) model. These methods provide approximations to the posterior distribution and are derived using the truncated stick-breaking representation and related approaches. We investigate their use in density estimation and cluster allocation and compare to Monte-Carlo results. Further, more specific topics include the general mathematical structure of the mean field approximation, the handling of the truncation level, the effect of including a prior on the concentration parameter α of the DPM model, the relationship between the proposed variants of the mean field approximation, and the connection to maximum a-posteriori estimation of the DPM model.

AMS 2000 subject classifications: Primary 62E17; secondary 62G07.
Keywords and phrases: Bayesian nonparametrics, approximation methods, variational inference, density estimation.

Received December 2008.

1. Introduction

Recently, variational approximation schemes have received growing interest as alternatives to Monte Carlo integration in computational Bayesian inference (Bishop 2006, Wainwright and Jordan 2008). In particular, mean field methods are now frequently being used in a variety of contexts (Opper and Saad 2001). In the mean field approach, an approximation to a complicated posterior distribution is found within a more tractable set of trial functions by means of minimizing the Kullback-Leibler distance. Under appropriate conditions, mean field methods can offer large computational gains combined with ease of application and adequate quantitative accuracy.

However, the mean field approach also has a number of drawbacks. It can only be used efficiently in certain classes of models, and it is often difficult to assess accuracy without comparison to exact results since generally applicable performance criteria are not known. Furthermore, in many cases it requires substantial effort to find tractable sets of trial functions that improve upon the commonly used fully factorized class.

Nevertheless, in many cases the advantages of the mean field method outweigh the associated problems, and it is one of the current topics of research in this area to identify such opportunities. Such work typically focuses on well-defined, specific classes of models since it appears to be rather difficult

to derive generally valid results. Very recently, a number of papers presented rather promising proposals regarding the application of mean field methods to inference in Dirichlet process mixture (DPM) models (Blei and Jordan 2006, Kurihara, Welling, and Teh 2007, Kurihara, Welling, and Vlassis 2007). In particular, using an image clustering problem as a practical example, Blei and Jordan (2006) showed that these methods are able to efficiently handle large high-dimensional data sets and may provide large computational gains compared to standard Markov chain Monte Carlo (MCMC) integration.

The purpose of this paper is to present a systematic study of these mean field approximation schemes. Several reasons motivate our interest in this problem. (i) The DPM model is very important as a generic building block for nonparametric Bayesian inference (Ferguson 1973, Antoniak 1974). Computational inference for this model has been developed thoroughly over the past 15 years on the basis of MCMC integration (see, e.g., Escobar 1988; 1994, MacEachern 1994, Escobar and West 1995, MacEachern and Müller 1998, Neal 2000, Ishwaran and James 2001, Walker 2007). Nevertheless, in some situations, such as the image analysis problem mentioned above, it is still desirable to have complementary computational approaches available, so that detailed investigations of the mean field approximation are well justified. (ii) The previous studies of mean field inference for DPM models provide some assessment of the performance, e.g., by studying some large-scale classification problems or monitoring the dependence the Kullback-Leibler divergence on some parameters of the algorithm. However, a more systematic study of the basic properties of the method is still missing. It is also not clear how accurately mean field approximates the various posterior quantities calculated in the DPM model. In this work, we aim to address some of these issues. (iii) Mean field methods have already been applied to extensions of the DPM model (Teh, Kurihara, and Welling 2008) as well as to closely related mixture models such as latent Dirichlet allocation (Blei, Ng, and Jordan 2003). Some of the present results are expected to be of interest in these contexts as well. (iv) Since it constitutes a measure over probability measures, the Dirichlet process is quite different from the more standard ingredients of Bayesian probability models. It is therefore of interest to see how the mean field approximation behaves in this context.

The plan of the paper is as follows. In Sec. 2, we first summarize some basic facts about Dirichlet process mixtures. We then introduce the mean field approximation and outline its application to the DPM model in the truncated stick-breaking representation, as proposed by Blei and Jordan (2006). Section 3 discusses the general mathematical structure of the mean field solutions. The conclusions are applied to the problem of choosing the truncation level in the stick-breaking representation. We also discuss some instructive scenarios with one and two observations in more detail. Section 4 illustrates and extends the foregoing discussion by means of several representative numerical examples. We also compare to results obtained from MCMC integration. The comparison focuses on density estimation and cluster allocation. Several variants of the mean field approximation scheme that were proposed in Kurihara, Welling, and Teh (2007) are compared in Sec. 5. In Sec. 6 we study the mean field approximation

after inclusion of a prior for the DPM parameter α. The relationship between the mean field approximation and maximum a-posteriori (MAP) estimation of the posterior is discussed in Sec. 7. A summary and some concluding remarks are given in Sec. 8.

As a main result of our study we find that while the mean field methods, by themselves, provide a useful tool for mixture modelling, great care is necessary when using them to calculate actual approximations to the DPM posterior. In particular regarding the description of data clustering, they display great differences to the exact results.

2. Variational inference and the Dirichlet process mixture model

2.1. Dirichlet process mixtures

In the Dirichlet process mixture model (Ferguson 1973, Antoniak 1974), one assumes the observations to originate from a mixture distribution with an unknown number of components, and one places a Dirichlet process prior on the mixture distribution. More formally, the DPM model is defined by

$$\begin{aligned}(Y_i|\theta_i) &\sim p(Y_i|\theta_i), \quad i = 1, \ldots, n,\\ (\theta_i|G) &\sim G,\\ G &\sim DP(\alpha, G_0).\end{aligned}$$

Here, G denotes a random probability measure drawn from a Dirichlet process with base distribution G_0 and parameter α. The measure G is almost surely discrete. The ith observation Y_i is obtained from the distribution $p(Y_i|\theta_i)$, conditional on the parameter θ_i drawn from G. In this way, G can be thought of as representing a mixture distribution with an infinite number of components.

One of the main approaches to computing inferences for the DPM model makes use of the stick-breaking representation for the random measure G (Sethuraman 1994), i.e.,

$$G(\cdot) \stackrel{\mathcal{D}}{=} V_1\delta_{\eta_1}(\cdot) + \sum_{k=2}^{\infty} V_k \prod_{j=1}^{k-1}(1 - V_j)\delta_{\eta_k}(\cdot)$$

where the V_j's have a beta distribution $\mathcal{B}(1, \alpha)$, δ_η denotes the Dirac measure placing unit mass at η, and $\eta_j \sim G_0$. The measure G is thus described by the collection of random variables V_j and η_j. Ishwaran and James (2001) made the important observation that a truncation of the stick-breaking representation at a sufficiently large K, i.e., setting $v_K = 1$, already provides an excellent approximation to the full DPM model. For the truncated stick-breaking model, we can write down the explicit probability distribution function

$$\begin{aligned}&p(y_{1:n}, z_{1:n}, \eta_{1:K}, v_{1:K-1})\\ &\quad= \prod_{i=1}^{n}\prod_{k=1}^{K}[p(y_i|\eta_k)w_k(v_{1:K})]^{\mathbb{I}(z_i=k)} \prod_{k=1}^{K} G_0(\eta_k) \prod_{k=1}^{K-1}\mathcal{B}(v_k; 1, \alpha). \quad (1)\end{aligned}$$

Here, $y_{1:n}$, $z_{1:n}$, $\eta_{1:K}$ are shorthand for the sets of variables $y_1, \ldots, y_n$ etc., $v_{1:K-1} = v_{1:(K-1)}$, whereas the $w_k(v_{1:K}) := v_k \prod_{j=1}^{k-1}(1-v_j)$ denote the mixture weights. The indicator variables z_i describe the assignment of the ith observation to a specific mixture component and take integer values between 1 and K. Finally, $\mathbb{I}$ in (1) denotes the indicator function. In the form (1), the DPM is easily amenable to mean field variational methods.

The DPM model can be extended in various ways by placing priors on α and further parameters that may appear in the densities G_0 and $p(Y_i|\theta_i)$. However, to work out the properties of the mean field approximation most clearly, we will only consider the basic model (1) except for Sec. 6 where we discuss the effects of a prior distribution placed on the parameter α.

For later reference, we note that the marginal $p(y_{1:n})$ can be written as (Lo 1984, Lemma 2)

$$p(y_{1:n}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \sum_{c \in \mathcal{C}} f(c^{(1)}) \times \cdots \times f(c^{(m)}) \tag{2}$$

where the sum is over all possible partitions $c = (c^{(1)}, \ldots, c^{(m)})$ of the observations $y_{1:n}$, and $c^{(j)}$ denotes a particular subset (cluster) of observations. If $c^{(j)}$ contains the observations $y_1, \ldots, y_k$, say, the corresponding factor is given by

$$f(y_1, \ldots, y_k) = \alpha(k-1)! \int p(y_1|\eta) \times \cdots \times p(y_k|\eta) G_0(\eta) d\eta.$$

2.2. Mean field inference for the Dirichlet process mixture model

The basic idea of the mean field method is to approximate a complicated probability distribution p by a simpler one for which one can compute inferences more easily. To this end, one starts from a set $\mathcal{Q}$ of tractable trial functions for which both the mean field procedure as well as subsequent inference are feasible. The mean field approximation q is then found from the minimization of the Kullback-Leibler (KL) divergence between the functions in $\mathcal{Q}$ and p, i.e.,

$$q = \operatorname*{argmin}_{q' \in \mathcal{Q}} \int q' \log \frac{q'}{p}.$$

The quality of the result thus depends on the choice of the class $\mathcal{Q}$. Detailed discussions of the mean field method can be found, e.g., in Opper and Saad (2001), Bishop (2006), Wainwright and Jordan (2008).

To simplify notation in the subsequent discussions, we will denote all mean field trial functions and their constituent factors by the letter q. Any particular function or factor is then identified and distinguished by its arguments and through the context (and further subscripts, if necessary).

Mean field inference for the DPM model was first discussed by Blei and Jordan (2006) within the context of the stick-breaking representation, and we follow their approach with some slight modifications. In particular, while Blei and Jordan (2006) use the full DPM model as target distribution, we start from the

truncated posterior $p(z_{1:n}, \eta_{1:K}, v_{1:K-1}|y_{1:n})$ induced by (1) and then study how the mean-field approximation behaves in the limit of $K \to \infty$.

A tractable class $\mathcal{Q}$ is given by the set of all distributions factorizing as $q(z_{1:n})q(\eta_{1:K}, v_{1:K-1})$. To determine the actual mean field approximation, i.e., the minimizer of the KL distance, one can employ an iteration scheme where one alternatingly optimizes $q(z_{1:n})$ for a previously found, fixed $q(\eta_{1:K}, v_{1:K-1})$ (in the sense of minimizing the KL distance to p) and vice versa. More specifically, the iteration scheme proceeds as follows.

0. Obtain an initial guess for $q(\eta_{1:K}, v_{1:K-1})$ (alternatively, one might also start with a guess for $q(z_{1:n})$ and begin the iteration at step 2 below).
1. Update $q(z_{1:n})$ according to

$$q(z_{1:n}) \propto \exp\left[\mathbb{E}_{q(\eta_{1:K}, v_{1:K-1})}\left(\log p(y_{1:n}, z_{1:n}, \eta_{1:K}, v_{1:K-1})\right)\right]. \tag{3}$$

2. Update $q(\eta_{1:K}, v_{1:K-1})$ according to

$$q(\eta_{1:K}, v_{1:K-1}) \propto \exp\left[\mathbb{E}_{q(z_{1:n})}\left(\log p(y_{1:n}, z_{1:n}, \eta_{1:K}, v_{1:K-1})\right)\right]. \tag{4}$$

3. Calculate the negative free energy

$$\begin{aligned} H \quad = \quad & \int q(z_{1:n})q(\eta_{1:K}, v_{1:K-1}) \\ & \times \log \frac{p(y_{1:n}, z_{1:n}, \eta_{1:K}, v_{1:K-1})}{q(z_{1:n})q(\eta_{1:K}, v_{1:K-1})} dz_{1:n} d\eta_{1:K} dv_{1:K-1}. \end{aligned} \tag{5}$$

   This quantity provides a lower bound on the marginal $\log p(y_{1:n})$ and is guaranteed to increase in each iteration step (the latter fact also provides a useful check on numerics).
4. Repeat steps 1 to 3 until H no longer increases appreciably which signals the convergence of the iteration scheme.

A well-known problem of such mean-field iteration schemes is the existence of multiple fixed points corresponding to different stationary points of the KL divergence (MacKay 2003, Wainwright and Jordan 2008). Commonly, one restarts the iteration several times and selects the solution with the largest H as final mean-field approximation, but it is sometimes difficult in practice to find the global optimum (Weiss 2001).

From (1) and the above update relations it follows that the optimized distributions factorize in all variables. It is thus sufficient to perform the iteration with fully factorized functions [see (7), (9), (10) below]. A further important simplification arises if the base distribution $G_0(\eta)$ is chosen conjugate to the observation model $p(y|\eta)$ and both are in the exponential family. In this case, the mean-field approximation $q(\eta_k)$ will belong to the same class of distributions as G_0, and the iterations can be carried out by updating a finite vector of parameters. In the following, we will be using two related types of such models. For analytical study, we consider the example of a simple location model (Escobar 1994) which describes a mixture of normals with fixed variance. The

update equations adapted to this case are given below. In the numerical examples of the later sections, we will use a normal/inverse-gamma location-scale model (Escobar and West 1995) where the normal mixture components have variable variance. We expect that these models will also be very important in actual practical applications.

More specifically, the normal location model of Escobar (1994) is defined as

$$\begin{aligned} (Y_i|\theta_i) &\sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, \ldots, n, \\ (\theta_i|G) &\sim G, \\ G &\sim DP\left(\alpha, \mathcal{N}(0, \lambda^2)\right) \end{aligned} \tag{6}$$

with $\mathcal{N}$ the normal distribution and the variances λ^2 and σ^2 considered as fixed parameters. One can then derive the iteration relations

$$\begin{aligned} q(z_i = k) &\propto \exp\Bigg\{ \mathbb{E}_{\eta_k}[\log p(y_i|\eta_k)] \\ &\quad + \mathbb{E}_{v_k}(\log v_k) + \sum_{j=1}^{k-1} \mathbb{E}_{v_j}[\log(1 - v_j)] \Bigg\}, \; 1 \le k \le K, \end{aligned} \tag{7}$$

$$\propto \exp\left\{ -\frac{\rho_k^2 + (\mu_k - y_i)^2}{2\sigma^2} + \psi(n_k + 1) - \sum_{j=1}^{k-1} \frac{1}{\alpha + n_j^>} \right\}, k < K, \tag{8}$$

$$q(v_k) \sim \mathcal{B}(n_k + 1, n_k^> + \alpha), \tag{9}$$

$$q(\eta_k) \propto \exp\left[\sum_{i=1}^{n} \mathbb{E}_{z_i}(\mathbb{I}(z_i = k)) \log p(y_i|\eta_k) + \log G_0(\eta_k) \right], \tag{10}$$

$$\sim \mathcal{N}\left(\mu_k := \frac{\sum_i \pi_{ik} y_i}{\sigma^2/\lambda^2 + \sum_i \pi_{ik}}, \quad \rho_k^2 := \frac{1}{1/\lambda^2 + \sum_i \pi_{ik}/\sigma^2} \right). \tag{11}$$

Here, we have introduced the abbreviation $\pi_{ik} := q(z_i = k)$ for the probability of assigning the ith observation to the kth mixture component which will be a key quantity in our subsequent discussion. We also define the mean occupation $n_k := \sum_{i=1}^n \pi_{ik}$ of the kth component, and set $n_k^> := \sum_{l=k+1}^K n_l$. Finally, $\psi(z) = \Gamma'(z)/\Gamma(z)$ denotes the digamma function.

Relations (7) and (9), (10), respectively, follow directly from (3) and (4) using fully factorized trial functions $\prod_i q(z_i)$ and $\prod_k q(v_k)q(\eta_k)$. They hold for arbitrary $p(y|\eta)$ and $G_0(\eta)$. As mentioned above, the fully factorized form does not impose any limitations compared to the more general case $q(z_{1:n})q(\eta_{1:K}, v_{1:K-1})$.

Relations (8) and (11) have been specialized to the normal location model (6). To simplify notation, we have omitted an explicit index for counting iteration cycles. To derive (8), we have made use of the relation $\psi(z + 1) - \psi(z) = 1/z$. Note that in relation (8), the right-hand side values of n_k and $n_k^>$ are calculated from the previous iteration step, i.e., (8) is not an implicit equation for the π_{ik}'s. Furthermore, for $k = K$ the final two terms in the argument of the exponential

in (8) have to be replaced by $\psi(n_K+\alpha)-\sum_{j=1}^{K-2}(\alpha+n_j^{>})^{-1}$. Note that the π_{ik}'s succinctly characterize the converged solution since the functions $q(v_k)$ and $q(\eta_k)$ are straightforwardly derived from them.

The analytical expression for the lower bound (5) simplifies if it is calculated after the functions $q(v_k)$ have been updated using the current π_{ik}'s. In this case, a number of terms cancel and one obtains

$$\begin{aligned} H &= \sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}\mathbb{E}_{\eta_k}(\log p(y_i|\eta_k))+\sum_{k=1}^{K}\mathbb{E}_{\eta_k}(\log G_0(\eta_k))-\sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}\log\pi_{ik} \\ &\quad -\sum_{k=1}^{K}\mathbb{E}_{\eta_k}(\log q(\eta_k))+(K-1)\log\alpha+\log\frac{\Gamma(\alpha+n_K)}{\Gamma(\alpha+n)}+\sum_{k=1}^{K-1}\log\frac{\Gamma(n_k+1)}{n_k+n_k^{>}+\alpha}. \end{aligned} \tag{12}$$

Equation (12) also holds at convergence.

From the final mean-field approximation, we obtain information about clustering, i.e., the assignment of observations to different mixture components, through the function $q(z_{1:n})$. In addition, we can calculate the predictive distribution $p(y_{n+1}|y_{1:n})$ as (Blei and Jordan 2006)

$$p(y_{n+1}|y_{1:n})=\sum_{k=1}^{K}\mathbb{E}_{v_{1:K}}(w_k(v_{1:K}))\mathbb{E}_{\eta_k}(p(y_{n+1}|\eta_k)). \tag{13}$$

Note that (13) is normalized because of $\sum_{k=1}^{K}w_k(v_{1:K})=1$. For the normal location model

$$\mathbb{E}_{\eta_k}(p(y_{n+1}|\eta_k))=\frac{1}{\sqrt{2\pi(\sigma^2+\rho_k^2)}}\exp\left[-\frac{(y_{n+1}-\mu_k)^2}{2(\sigma^2+\rho_k^2)}\right]. \tag{14}$$

2.3. Mixtures of mean field approximations

It is well known that one can sometimes find improved mean-field approximations that lie outside the originally chosen class $\mathcal{Q}$ by considering mixtures of trial functions that have non-overlapping support (Mézard, Parisi, and Virasoro 1987). For our purposes, a precise formulation of this statement can be given as follows.

Proposition. Let $p(x)$ be the target distribution and $q_1(x),\ldots,q_m(x)$ a set of functions from $\mathcal{Q}$ with respective KL distances $K_1,\ldots,K_m>0$ to p. It is assumed that the functions q_i have non-overlapping support, i.e., $q_i(x)>0$ for a given x implies that $q_j(x)=0$ for all $j\neq i$. The best approximation of the target distribution $p(x)$ within the class of mixtures $\sum_{i=1}^{m}w_iq_i(x)$ is given by

$$q(x)=\sum_{i=1}^{m}\frac{\exp(-K_i)}{\sum_{k=1}^{m}\exp(-K_k)}q_i(x). \tag{15}$$

Its KL distance $-\log[\sum_{k=1}^{m}\exp(-K_k)]$ is strictly smaller than all individual distances K_i so that (15) always yields an improved approximation.

Proof. Setting $q'(x) := \sum_{i=1}^m w_i q_i(x)$, it follows that $\int dx\, q'(x) \log q'(x) = \sum_i w_i \int dx\, q_i(x) \log q_i(x) + \sum_i w_i \log w_i$ since the functions $q_i(x)$ have non-overlapping support. The KL distance between q' and p is therefore given by $K_0 = \sum_i w_i K_i + \sum_i w_i \log w_i$. Minimization of K_0 with respect to the w_i under the constraint $\sum_i w_i = 1$ leads to (15) and the given expression for the minimum KL distance. □

In the discussion of (15), the following points should be mentioned. (i) The condition of non-overlapping support is crucial since it allows to explicitly evaluate the entropy term $\int dx\, q'(x) \log q'(x)$ as given above. Otherwise, the calculation of the entropy becomes intractable and one has to resort to approximation schemes as described, e.g., in Jaakkola and Jordan (1998). (ii) The functions $q_1, \ldots, q_m$ may be chosen arbitrarily within the class $\mathcal{Q}$ as long as they fulfill the support condition, and need not be variational extrema. (iii) Since, in general, the mixture will not be a member of the original class $\mathcal{Q}$ of trial functions one can indeed obtain improved approximations that go beyond $\mathcal{Q}$. (iv) The lower bound defined in (5) is related to the KL distance by $H = -K + \log p(y_{1:n})$. Equation (15) thus remains valid with the $K_{i(k)}$ replaced by $-H_{i(k)}$, and the improved lower bound is given by $\log[\sum_{k=1}^m \exp(H_k)]$. (v) As an interesting aside, the fact that the minimized KL distance has to be non-negative implies that $\sum_{i=1}^m \exp(-K_i) \leq 1$ for any choice of the non-overlapping distributions q_i. In the extreme case of one q_i coinciding with p, all the other KL distances K_j, $j \neq i$, equal $-\infty$.

In the context of the DPM model, we will see that the distributions corresponding to the various fixed points of the mean field iteration scheme often come very close to fulfilling the condition of non-overlapping support. This suggests to construct improved mean-field approximations by using these distributions in a mixture. Thereby, one has to neglect the small error introduced by the remaining overlap. Alternatively, one can choose functions $q_1, \ldots, q_m$ such that they strictly fulfill the overlap condition and are still close to the fixed-point distributions.

3. Mean field inference for the DPM model: General features

In this section, we give an overview and explanation of some of the main distinctive features that characterize the mean field approximation to the posterior in the DPM model. For clarity, the discussion will be based on the normal location model defined in (6). As demonstrated by the numerical studies of Sec. 4, the essential conclusions also apply to the more general location scale model of Escobar and West (1995). Throughout the following discussion, we will presuppose a vague prior distribution, i.e., $\lambda \gg \sigma$. This condition is not only reasonable from a modelling point of view since it guarantees sufficient flexibility in describing the mixture distribution, but it also greatly simplifies the behavior of the mean field approximation.

In the following, we will first discuss the MF approximation for a single observation. The detailed study of this simple case which can largely be performed

analytically is very instructive since it already reveals several important features that can also be found in more complex problems. We next turn to the case of multiple observations and discuss several aspects of the general mathematical structure of the mean field approximations. Our findings are then applied to discuss the problem of choosing the truncation level K. Finally, we return to a simple model problem with two data points in order to illustrate how the mean field approximation switches between different clusterings of data points.

Let us start, however, by briefly reviewing some basic features of the DPM mean-field approximation and the update relations (8)–(11). The mean-field treatment is based on a truncation of the DPM model. The truncated model describes the mixture in terms of a finite number K of components, in contrast to the infinite mixture of the full problem. If K is chosen sufficiently large, the truncation provides an excellent approximation to the full model (Ishwaran and James 2001).

After convergence of the iteration procedure, the functions $q(\eta_k)$ and $q(v_k)$ yield the mean-field approximations to the posterior marginal distributions of the location and weight parameters η_k and v_k, respectively, for the kth mixture component. The functions $q(z_i)$ describe the assignment probability of the ith observation to the different components. As we see from (9) and (11), the more observations are attributed to a particular component, the more sharply the posteriors $q(\eta_k)$ and $q(v_k)$ will be localized. The assignment probabilities $\pi_{ik} = q(z_i = k)$ are determined by two factors. First of all, we have the prior contribution $\exp\left[\psi(n_k + 1) - \sum_{j=1}^{k-1}(\alpha + n_j^{>})^{-1}\right]$. This factor is closely related to the last term of (12). As follows from the discussion in point (iii) of Sec. 3.2, it ensures that the optimal mean field approximation (i.e., the iteration fixed point with largest H) populates mixture components with small k.

The second and more interesting contribution is derived from the data model $p(y|\eta)$. It shows that an observation y_i will tend to be assigned to components for which the location posterior $q(\eta_k)$ overlaps with y_i (term $\exp[-(\mu_k - y_i)^2/2\sigma^2]$) and, crucially, which are well localized (term $\exp(-\rho_k^2/2\sigma^2)$). Since the localization depends on the number of observations assigned to a component, the latter aspect leads to a "positive feedback" or "self-reinforcement" process that results in individual observations being assigned to a small number of components, often even only a single one.

3.1. Single observation

As mentioned above, in the case of a single observation ($n = 1$), the mean field approximation is amenable to a detailed analytical study. The results obtained are useful since they already anticipate some important traits of the general case. In the following discussion, we will set $y_1 = 0$ for simplicity.

At first, let us briefly discuss how the self-reinforcement effect described above manifests itself in this case. From (8)–(11), it follows that the relevant relations

for the iteration scheme are given by

$$\pi_{1k} \propto \exp\left\{-\frac{\rho_k^2}{2\sigma^2} + \psi(\pi_{1k}) - \sum_{j=1}^{k-1} \frac{1}{\alpha + \pi_{1,>j}}\right\} \tag{16}$$

with $\rho_k^2 = (1/\lambda^2 + \pi_{1k}/\sigma^2)^{-1}$ (the relation for π_{1K} is slightly modified). Suppose that we start the iteration procedure from the initial condition $\pi_{1k} = 1/K$. After the first iteration cycle, all ρ_k^2 will be equal, but due to the terms $-\sum_{j=1}^{k-1}(\alpha + \pi_{1,>j})^{-1}$ in the exponent of (16), the assignment probabilities π_{1k} will be ordered as $\pi_{11} > \pi_{12} > \cdots > \pi_{1K}$, in general. In the next iteration cycle, the ρ_k^2's thus will also become ordered in this way which causes π_{11} to grow even further at the expense of the other components. Subsequent cycles will rapidly enhance this process, until the observation is almost completely assigned to the first component. The exponential function in relation (16) is crucial in bringing about this behavior since it "magnifies" the effect of any differences of its arguments for different k. To see that our description is consistent, we note that after convergence, we will have $\rho_1^2 \approx \sigma^2$, whereas $\rho_k^2 \approx \lambda^2$, $k > 1$. This implies that the population of all components $k > 1$ will indeed be suppressed exponentially by a factor of $\xi = \exp(-\lambda^2/2\sigma^2)$ since we have assumed $\lambda \gg \sigma$.

Further studies reveal that $\pi_{1k} \approx \mathbb{I}(z_1 = 1)$ is not the only possible fixed point of the iteration scheme, but there are also solutions of the form $\pi_{1k} \approx \mathbb{I}(z_1 = l)$, $l > 1$, which arise if the initial conditions are chosen appropriately. Approximate analytical expressions for these solutions can be derived in terms of an expansion in ξ. From relations (8) and (11) it follows that the converged assignment probabilities π_{1k} obey the equations

$$\pi_{1k} = \frac{\exp\left\{-\frac{1}{2(\sigma^2/\lambda^2+\pi_{1k})} + \psi(\pi_{1k}+1) - \sum_{j=1}^{k-1}\frac{1}{\alpha+\pi_{1,>j}}\right\}}{\sum_{l=1}^{K-1}\exp\left\{-\frac{1}{2(\sigma^2/\lambda^2+\pi_{1l})} + \psi(\pi_{1l}+1) - \sum_{j=1}^{l-1}\frac{1}{\alpha+\pi_{1,>j}}\right\} + e_K}, \tag{17}$$

$k < K$, with $e_K = \exp\{-[2(\sigma^2/\lambda^2+\pi_{1K})]^{-1} + \psi(\pi_{1K}+\alpha) - \sum_{j=1}^{K-2}(\alpha+\pi_{1,>j})^{-1}\}$. We can now seek solutions of the form $\pi_{1l} = 1-\delta_l$, $\pi_{1k} = \delta_k$, $k \neq l$, with $\delta_j \ll 1$ for all j. These are found to be given by

$$\begin{aligned}
\pi_{1l} &\approx 1 - \exp\left[-\frac{\lambda^2}{2\sigma^2} + \frac{1}{2\left(1+\frac{\sigma^2}{\lambda^2}\right)} - 1\right] \\
&\quad \times \left[\frac{\exp\left(\frac{l-1}{\alpha+1}\right)-1}{1-\exp\left(-\frac{1}{1+\alpha}\right)} + e^{-1/\alpha}\frac{1-\exp\left(-\frac{K-l}{\alpha}\right)}{1-\exp(-1/\alpha)}\right] - \pi_{1K}, \qquad (18)\\
\pi_{1k} &\approx \exp\left[-\frac{\lambda^2}{2\sigma^2} + \frac{1}{2\left(1+\frac{\sigma^2}{\lambda^2}\right)} - 1\right]\exp\left(\frac{l-k}{\alpha+1}\right), \quad k < l, \qquad (19)\\
\pi_{1k} &\approx \exp\left[-\frac{\lambda^2}{2\sigma^2} + \frac{1}{2\left(1+\frac{\sigma^2}{\lambda^2}\right)} - 1\right]\exp\left(-\frac{k-l}{\alpha}\right), \quad l < k < K, \qquad (20)
\end{aligned}$$

$$\pi_{1K} \approx \exp\left[-\frac{\lambda^2}{2\sigma^2}+\frac{1}{2\left(1+\frac{\sigma^2}{\lambda^2}\right)}+\psi(\alpha)-\psi(2)\right]\exp\left(-\frac{K-2-l}{\alpha}\right), \quad (21)$$

where the neglected subsequent terms in the expansion are at least of order ξ^2 including logarithmic corrections. Since we have assumed that $\xi \ll 1$, we see that the solutions (18)–(21) indeed have the desired structure, since all correction terms δ_j remain small. For $l = K$ the expressions are slightly modified. We will return to some features of the solutions in Sec. 3.2.

The existence of multiple fixed points is a generic feature of the mean-field method. They are all local maxima, or at least stationary points, of the lower bound H. In the present case, (12) shows that the value of H decreases with growing l so that one would choose the solution for $l = 1$ as final mean-field result. We also note that the above expansion breaks down for growing l, as the terms $\exp[(l-k)/(\alpha+1)]$ will eventually become comparable to ξ^{-1}. However, since this typically only happens for large l, this complication is not of practical relevance. We also remark that the above derivation does not preclude the existence of other types of mean-field solutions with a different mathematical structure, but none could be found in numerical investigations.

An interesting and important aspect of relations (18)–(21) concerns the parameter α. One sees that the essential structure of the solution, i.e., the compression of the observation into a single component, is not at all affected by α. The choice of α thus is hardly of relevance for the posterior distribution. It is instructive to compare this behavior to the exact solution for the DPM model. From (1), one obtains the posterior marginal

$$p(z_1 = k|y_1) = \frac{1}{\alpha}\left(\frac{\alpha}{\alpha+1}\right)^k, \quad (22)$$

for $K \to \infty$ (and any value of y_1), which is of completely different character than the mean-field result. The distribution only depends on α, but not on λ or σ, and the observation is typically spread over several components. The mean-field solution thus does not yield a good description of this posterior marginal.

On the other hand, from (13) and the mean field solution for $l = 1$, we obtain the predictive density

$$\begin{aligned} p(y_2|y_1=0) \approx\ & \frac{\kappa}{\kappa+\alpha}\sqrt{\frac{1+\sigma^2/\lambda^2}{2\pi\sigma^2(2+\sigma^2/\lambda^2)}}\exp\left(-\frac{y_2^2(1+\sigma^2/\lambda^2)}{2\sigma^2(2+\sigma^2/\lambda^2)}\right) \\ & + \frac{\alpha}{\kappa+\alpha}\frac{1}{\sqrt{2\pi(\sigma^2+\lambda^2)}}\exp\left(-\frac{y_2^2}{2(\sigma^2+\lambda^2)}\right) \end{aligned} \quad (23)$$

with $\kappa = 2$. This result is obtained after slightly simplifying the mean field solution by setting $\pi_{1k} = \mathbb{I}(k = 1)$ and using the expressions for $q(v_k)$ and $q(\eta_k)$ ensuing from a single iteration of the update equations. The exact relation for the predictive density can be derived from (2) and is given by expression (23) with $\kappa = 1$. Apart from this small difference in the weight factors of the Gaussians, the mean field and exact predictive densities are identical.

The foregoing discussion already indicates a general trend that is observed in numerical studies of more complex problems. The mean field results often provide a reasonable approximation to the predictive density, whereas there are strong deviations for cluster allocation, which is determined by the posterior marginals for $z_{1:n}$.

Interestingly, for a single observation it is possible to construct a closer relationship between the mean field approximation and the exact description. To this end, we consider a mixture of trial functions from $\mathcal{Q}$ as discussed in Sec. 2.3. We choose the distributions q_m, $m = 1, 2, \ldots$, such that $q_m(z_i) = \mathbb{I}(z_i = m)$ whereas $q_m(v_k)$ and $q_m(\eta_k)$ are obtained from a single iteration of (9) and (11), i.e., $q_m(v_k) = \mathcal{B}(v_k; 1, \alpha + 1)$ for $k < m$ etc. The functions q_m are non-overlapping, and they provide very good approximations of the actual mean-field solutions described above as long as m does not become too large. If we now calculate the bounds H_m and construct the mixture q_{mix}, we find that $q_{\mathrm{mix}}(z_1)$ coincides with the exact marginal (22) of the full model.

The above discussion has shown that for a single observation one can systematically improve the mean-field approximation by considering mixtures of different fixed-point solutions. However, further investigations indicate that for multiple observations, one cannot expect to approach the exact distribution in this way, in general. Moreover, the rapidly growing number of necessary distributions would make this method impractical. Nevertheless, one should keep in mind that the combination of non-overlapping mean-field solutions will always provide some improvement to the individual distributions and might therefore be taken into consideration under appropriate conditions.

3.2. Multiple observations

For multiple observations, the mean field approximation can only be studied numerically, in general. Nevertheless, typical solutions share several important structural properties that are already apparent in the special cases of relations (18)–(21). A more formal explanation of this behavior is given in the Appendix.

(i) In the converged mean-field solutions, there is a clear distinction between occupied and unoccupied mixture components. The population of unoccupied components is exponentially small, i.e., of order $\mathcal{O}(\exp(-\lambda^2/2\sigma^2))$, and thus goes to zero as $\lambda/\sigma \to \infty$. The population of occupied components remains finite in this limit. This allows to unambiguously identify occupied and unoccupied components. Note, however, that typically there will be a large number of distinct mean field solutions that differ in the assignment probabilities of the occupied components (as exemplified, e.g., in (18)–(21) by the choice of different l's).

(ii) For fixed λ/σ, the assignment probabilities π_{ik} do not change appreciably if the truncation level K is varied. For example, for a single observation we see explicitly from relations (18)–(21) that K provides only an exponentially small correction to the occupied component, whereas the empty ones do not depend on K in leading order. Furthermore, as indicated by (20), the populations of the

empty components following the last occupied one scale with k as $\exp(-k/\alpha)$ at fixed λ/σ.

(iii) As long as the influence of the terms $\sum_{j=1}^{k-1}(\alpha + n_j^>)^{-1}$ in the iteration equations remains small, permutation of indices in a mean field solution leads to another (approximately) valid solution. The last term of (12) then shows that the lower bound H is maximized if the occupied components are assigned the lowest indices, in the order $n_1 \geq n_2 \geq \ldots$. This property can usefully be exploited in cluster relabelling algorithms that accelerate the determination of the optimal mean field approximation (Kurihara, Welling, and Teh 2007).

An interpretation of the separation into occupied and empty components can be given as follows. The exact predictive distribution can be decomposed into a part where the new observation y_{n+1} is assigned to one of the pre-existing clusters and a further contribution for which y_{n+1} is placed into a new cluster of its own (see (2)). The functional form of this further contribution is given by $p_{\text{new}}(y_{n+1}) = \int p(y_{n+1}|\theta)p(\theta)d\theta$ and it has an overall relative weight of $\alpha/(\alpha + n)$. As $\lambda \gg \sigma$, the width of $p_{\text{new}}(y_{n+1})$ is of the order of λ. In plots of the predictive density for sufficiently small n, the contribution thus shows up as a broad unstructured "background" onto which the more structured and localized parts derived from the observations $y_{1:n}$ are superimposed.

If we assume a mean field solution for which the first k_0 components are occupied, it follows from (13) and (14) that the unoccupied components provide a total contribution of

$$p_{\text{empty}}(y_{n+1}|y_{1:n}) = \frac{\alpha}{\alpha + n + 1} \prod_{k=1}^{k_0-1} \frac{\alpha + N_{>k}}{1 + \alpha + N_{>k}} p_{\text{new}}(y_{n+1}) \tag{24}$$

to the predictive density. In this way, we can associate the empty components with the contribution $\alpha/(\alpha + n)p_{\text{new}}(y_{n+1})$ in the exact model in which y_{n+1} is placed in its own cluster. Furthermore, we see that the mean field term is always reduced in weight in comparison to the full DPM model, but in practice, the two contributions are often found to be quite similar in magnitude.

3.3. The role of the truncation level K

The structural properties of the mean field solutions summarized in Sec. 3.2 are already very helpful in the discussion of the choice of the truncation level K. This issue is an important question in numerical calculations. In previous work (Blei and Jordan 2006, Kurihara, Welling, and Vlassis 2007), K was considered as an additional variational parameter, and various schemes for optimally choosing K were proposed on the basis of maximizing a variational lower bound. However, since the truncated DPM model tends towards the full DPM model as $K \to \infty$ (Ishwaran and Zarepour 2002), one may wonder if the mean field approximation introduced in Sec. 2.2 also approaches some well-defined limiting behavior.

That this is indeed the case is already implied by the discussion of Sec. 3.2. First of all, the assignment probabilities π_{ik} of occupied components that deter-

mine all interesting properties of a particular mean field solution are insensitive to the choice of K. Second, the lower bound H for any specific solution rapidly approaches a well-defined limit for $K \to \infty$. To see this we note that a component which is completely empty (i.e., $\pi_{ik} = 0$ for all i) does not contribute at all to H. The combined contribution of all unoccupied components is thus of order $\exp(-\lambda^2/\sigma^2)$ and remains bounded as $K \to \infty$ because of the $\exp(-k/\alpha)$ scaling. As a consequence, we conclude that it is not necessary to find an optimized value for K. Rather, one can eliminate this additional degree of freedom in a systematic way by considering the limit of $K \to \infty$.

In fact, for practical purposes it is even unnecessary to calculate the asymptotic lower bounds precisely, since the differences in H between the various extrema typically are sufficiently large so that the solutions can be ranked unambiguously even with some remaining uncertainty in H. In practice, it often suffices to choose K just somewhat larger than the maximum number of occupied modes. We also note that the predictive density is insensitive to the choice of K as well.

3.4. Two observations

The case of two observations, $n = 2$, provides a simple way of obtaining some insight into how the mean field method switches between different clusterings of observations. Two observations can be assigned to either a single or two different clusters. In view of the results of Secs. 3.1, we expect that the first situation corresponds to a mean-field solution where $q(z_i) \approx \mathbb{I}(z_i = 1)$, $i = 1, 2$, whereas for the second one we have $q(z_1) \approx \mathbb{I}(z_1 = 1)$, $q(z_2) \approx \mathbb{I}(z_2 = 2)$ or vice versa. Numerical studies indeed show that a converged solution close to the first type always exists, while the second type exists as long as the two observations are sufficiently far away (on the scale set by σ). In the spirit of the mean field approach, the type of clustering is determined by the solution with the larger lower bound H. To obtain some quantitative understanding of the behavior of the mean field approximation, we assume the data to be given by $y_1 = y$ and $y_2 = -y$. Using (12), we then calculate the lower bounds after approximating the two mean-field solutions by $q(z_i) = \mathbb{I}(z_i = 1)$ and $q(z_1) = \mathbb{I}(z_1 = 1)$, $q(z_2) = \mathbb{I}(z_2 = 2)$, respectively. The functions $q(v_k)$ and $q(\eta_k)$ are obtained from a single iteration of (9) and (11). The lower bounds that are calculated analytically in this way are very close to the ones obtained numerically from the actual mean field distributions. In particular, one finds that the two bounds become equal for

$$\frac{y^2}{\sigma^2} = \left(1 + \frac{\sigma^2}{\lambda^2}\right)\left[\log\left(\frac{\lambda}{\sigma}\frac{1+\frac{\sigma^2}{\lambda^2}}{\sqrt{2+\frac{\sigma^2}{\lambda^2}}}\right) - \log\frac{\alpha}{2(\alpha+1)}\right]. \tag{25}$$

This means that the mean-field approximation switches abruptly between assigning the observations to a single or to two clusters, respectively, when y

crosses a threshold that is of the order of $\sigma \log(\lambda/\sigma)$. We also note that unless α is very small, the dependence on α is very weak.

In the full DPM model, one can infer from (2) that the probabilities for having one or two clusters change continuously upon varying y, i.e., there is a gradual transition in the cluster assignment. In particular, the two probabilities become equal for

$$\frac{y^2}{\sigma^2} = \left(1 + \frac{\sigma^2}{\lambda^2}\right) \left[\log\left(\frac{\lambda}{\sigma} \frac{1 + \frac{\sigma^2}{\lambda^2}}{\sqrt{2 + \frac{\sigma^2}{\lambda^2}}}\right) - \log\alpha\right]. \tag{26}$$

We note that those parts of (25) and (26) that only depend on the data likelihood parameters σ and λ are identical, whereas there are strong discrepancies in the dependence on the prior as expressed through α. We can interpret this result in the following way. Overall, (25) indicates that the mean field method displays a plausible behavior regarding clustering. One would expect that there should be two clusters as soon as y becomes comparable to σ. A similar result is also given by the full DPM model. However, the detailed comparison between the mean field and the exact treatment shows clear quantitiatve and qualitative differences, so that it is hard to consider the mean field calculation as an "approximation" to the full model. These overall conclusions are confirmed by the numerical study of more complicated situations.

In the present case, one can improve the mean-field result by considering mixtures of the individual mean field solutions as in Sec. 3.1. However, as mentioned above, solutions assigning the observations to different clusters do not exist if y is small compared to σ so that this approach is not always applicable. We also note that numerically we have not been able to find mean field solutions that make partial assignments of the observation to the mixture components (i.e., both π_{11} and π_{12} being large simultaneously, for example). In Sec. 4, we will discuss a numerical example of two partially overlapping clusters of observations that shows a very similar behavior to the simple case discussed here (see Fig. 2).

4. Numerical examples

In this section we will discuss a number of representative numerical examples that further illustrate the behavior of the mean field approximation beyond the general discussion of Sec. 3. The purpose of this discussion is, on the one hand, to show that the mean field method, when viewed on its own, provides "reasonable" results regarding clustering and density estimation. On the other hand, we want to point out the profound differences to the exact treatment of the DPM model, in particular regarding data clustering.

All calculations use the normal/inverse-gamma location-scale model of Escobar and West (1995) that was briefly introduced in Sec. 2.2. Formally, the

model is defined as

$$\begin{aligned}(Y_i|\mu_i,\nu_i) &\sim \mathcal{N}(\mu_i,\nu_i), \quad i=1,\ldots,n,\\ (\mu_i,\nu_i|G) &\sim G,\\ G &\sim DP\,(\alpha,G_0)\,,\end{aligned} \tag{27}$$

where the base distribution G_0 has density

$$\begin{aligned}g(\mu,\nu) &= g_{\mu|\nu}(\mu|\nu)g_\nu(\nu)\\ &= \mathcal{N}(\mu;0,\tau\nu)\mathcal{IG}(\nu;s/2,(T/2)^{-1}).\end{aligned} \tag{28}$$

Here, $\mathcal{IG}$ denotes the inverse gamma distribution with density $p(x;\alpha,\beta) = \beta^\alpha x^{-\alpha-1}\exp(-\beta/x)/\Gamma(\alpha)$ and shape and scale parameters α and β. The parameters s, T, and τ (as well as α) are considered fixed. Note that in the limit of $s,T\to\infty$, this model reduces to the location model (6) with $\sigma^2 = T/s$ and $\lambda^2 = \tau T/s$. It is therefore convenient to reparametrize the above model in terms of $\sigma^2_{\text{eff}} = T/s$ and $\lambda^2_{\text{eff}} = \tau T/s$, so that s provides a measure of the deviation from the location model with the corresponding parameters. In the truncated stick-breaking approximation to the full model (27), one obtains the probability distribution

$$\begin{aligned}&p(y_{1:n},z_{1:n},\eta_{1:K},V_{1:K},v_{1:K-1})\\ &= \prod_{i=1}^{n}\prod_{k=1}^{K}[p(y_i|\eta_k,V_k)w_k(v_{1:K-1})]^{\mathbb{I}(z_i=k)}\prod_{k=1}^{K}p(\eta_k|V_k)p(V_k)\prod_{k=1}^{K-1}\mathcal{B}(v_k;1,\alpha),\end{aligned}$$

where $p(y|\eta,V)$ is a normal distribution with mean η and variance V, $p(\eta|V)$ is normal with mean 0 and variance τV, and $p(V)$ is given by the inverse gamma distribution defined in (28). The mean-field update equations are calculated using relations (3) and (4), but we do not reproduce them here as the explicit expressions will not be needed in the following.

In the first example we consider three well separated clusters of observations. Each cluster consists of 30 data points drawn from the normal distributions $\mathcal{N}(-0.5,0.1^2)$, $\mathcal{N}(0,0.01^2)$, and $\mathcal{N}(0.5,0.04^2)$, respectively. The prior parameters are chosen as $\sigma_{\text{eff}} = 0.04$, $\lambda_{\text{eff}} = 1$, and shape parameter $s = 1$. The parameters have been selected such that the cluster widths are well within the support of the prior distribution for the variance of the mixture components.

In the mean field posterior, all data points within a cluster are assigned to the same mixture component, i.e., the approximation predicts exactly three components. This behavior is independent of α within the investigated range of $\alpha = 1$ to $\alpha = 50$. The predictive distribution for $\alpha = 1$ is shown in Fig. 1 (black curve) together with the result of an MCMC calculation based on algorithm 8 of Neal (2000). Both results are in very good agreement. This indicates that the posterior distributions of the mixture parameters are well localized. As shown in the inset of Fig. 1, however, there are significant differences between MCMC and mean field regarding the distribution for the number of mixture components. As mentioned, mean field predicts three components for all values of α. The MCMC

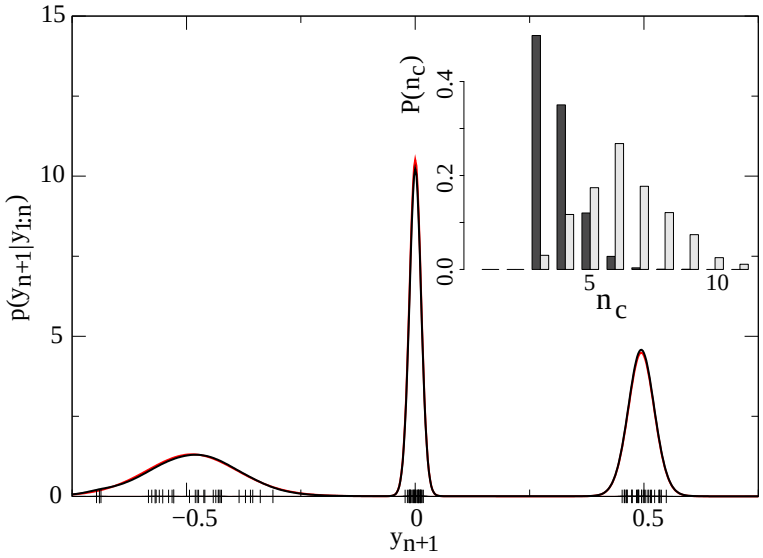

FIG 1. *Predictive density* $p(y_{n+1}|y_{1:n})$ *calculated for* $n = 90$ *observations indicated on the horizontal axis. The DPM parameters are* $\sigma_{\text{eff}} = 0.04$, $\lambda_{\text{eff}} = 1$, $s = 1$, $\alpha = 1$. *Black curve: mean field result, red: MCMC integration. Inset shows posterior* $P(n_c)$ *for number of mixture components, obtained from MCMC integration, for* $\alpha = 1$ *(dark bars) and* $\alpha = 5$ *(light). MF would yield a single bar at* $n_c = 3$.

result, which we can expect to be quite close to the exact distribution in the DPM model, extends over several numbers of components and depends strongly on α. We also note that although the mean field iteration scheme has a number of different fixed point solutions, none of the ones found numerically has more than 4 components. It would thus not be possible to recover the exact component distribution by mixing different mean field solutions. The compression into a very small number of components can be regarded as a consequence of the "positive feedback" mechanism described at the beginning of Sec. 3.

Besides making the differences to exact results for the DPM model obvious, this example also illustrates that the mean field scheme is indeed able to distinguish between clearly separated clusters of observations. This behavior, which could be regarded as a minimum requirement for any clustering method, has been confirmed in further numerical studies, and it can also be explained directly from the mean field iteration scheme. For the further discussion, it is thus of interest to see how the mean field methods handles more intricate situations in which one would expect partial assignment of observations to more than one component.

In our second example, we therefore consider a case where the data points originate from a mixture consisting of two partially overlapping normal distributions. To study this problem more systematically, we generate two sets of observations $\{y_i^{(l)}\}$, $l = 1, 2$, that each contain 40 data points drawn from a

$\mathcal{N}(0, 0.1^2)$ distribution. We then apply shifts of Δy and $-\Delta y$, respectively, to the points in the two sets. From our previous discussion, it is clear that for the two cases $\Delta y \ll \rho$ and $\Delta y \gg \rho$, the mean field approximation will firmly assign the data to a single and to two different mixture components, respectively. For the numerical study of the intermediate case, the parameters of the DPM model are set to $\alpha = 5$, $\sigma_{\text{eff}} = 0.1$, $\lambda_{\text{eff}} = 1$, and $s = 5$. With this relatively large value of s, the variance of the mixture components in the DPM model is fixed quite strongly.

The mean field approximation shows a behavior that displays some parallels to the case of two observations discussed in Sec. 3.4. For all investigated values of Δy, there is a mean-field fixed-point solution where all observations are assigned to the same mixture component. For large Δy, however, the optimal mean field solution, that has the largest lower bound H, populates two different components. As shown in Fig. 2(a), the lower bounds strongly depend on Δy, and upon decreasing Δy, the one-component solution eventually becomes dominant. Interestingly, the point at which the two bounds become equal is very well estimated by the simple relation (25). Soon after this crossover, the two-component solution becomes unstable and ceases to exist.

Figures 2(b) and (c) display some features of the mean field solution for $\Delta y = 0.15$ where the two-component solution is still slightly dominant. Figure 2(b) shows the assignment probabilities π_{ik}, $k = 1, 2$, for the two occupied mixture components. They are essentially determined by relation (8) if we disregard the prior variability of σ^2. The figure demonstrates that the mean field solution indeed makes partial assignments of the central data points to the two mixture components. The predictive distribution is shown in Fig. 2(c) (black curve) together with the result from the subdominant one-component solution (blue) and the MCMC calculation (red). The two insets show the predictive density for $\Delta y = 0.1$ and $\Delta y = 0.18$ where there is strong dominance of the one- or the two-component mean field solution, respectively. In each case, the difference between MCMC and mean-field result is somewhat larger than in the example of Fig. 1. We also note that a simple mixing of the two mean-field solutions at $\Delta y = 0.15$ makes the two-peak structure disappear and therefore does not produce an obvious improvement.

In order to illustrate the mean field method with a "realistic" data set, we have studied the well-known galaxy red shift data discussed in Roeder (1990). Figure 3 displays the results of corresponding calculations for parameters $\alpha = 1$, $\sigma_{\text{eff}} = 0.707$, $\lambda_{\text{eff}} = 7.07$, and $s = 4$. These values were chosen based on the discussion by Escobar and West (1995) who, however, also include hyperpriors for some of the parameters. The optimal mean field approximation (in terms of the lower bound H) that could be found numerically contains three mixture components (black curve in Fig. 3) whereas the second-best has four components (blue). In contrast, in the MCMC calculation for the DPM model (red), the posterior distribution for the number of components is spread out between 6 and 9. One can see from Fig. 3 that mean field and MCMC virtually agree in the description of the predictive density for the two well-separated outer clusters at low and high velocities. However, some discrepancies arise for the central part

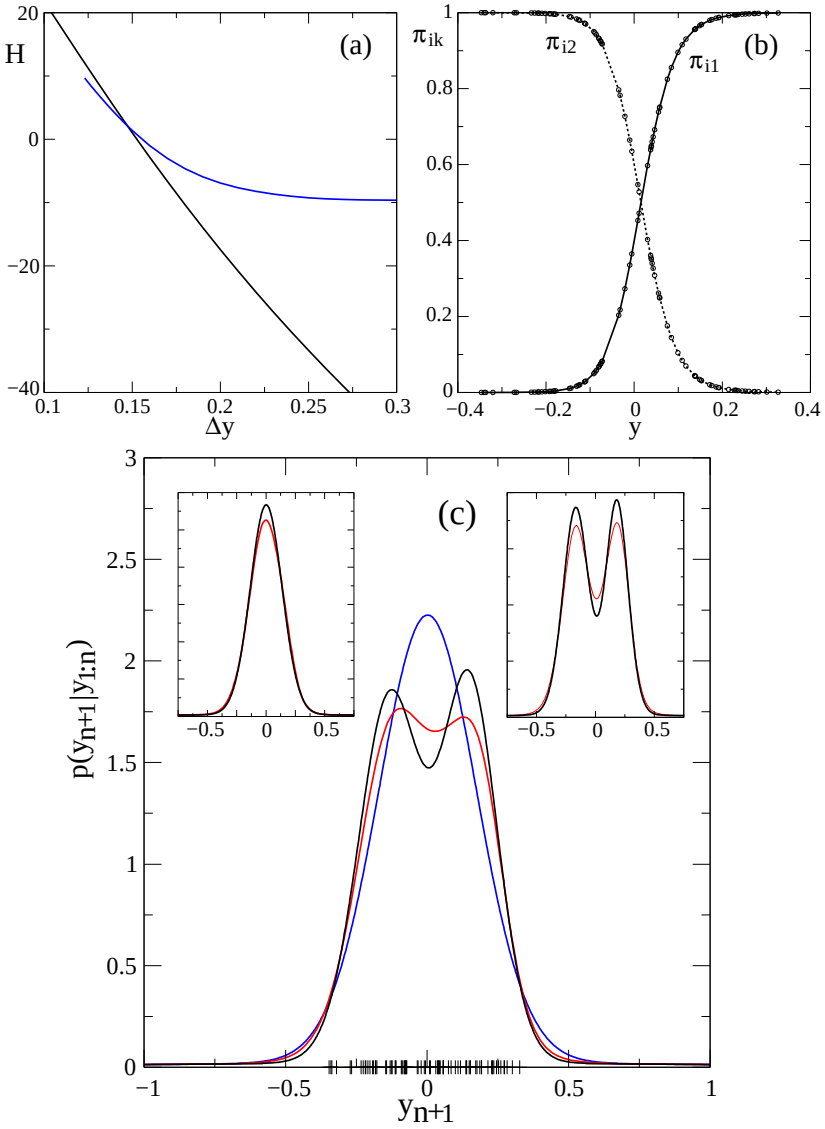

FIG 2. *Mean field approximation for data drawn from a two-component mixture. (a) Lower bounds H of one- and two-component mean field solutions (black and blue curve, respectively) as a function of the shift Δy applied to the mixture components. (b) Assignment probabilities π_{ik} for two-component mean field solution at $\Delta y = 0.15$. (c) Predictive densities at $\Delta y = 0.15$, $\Delta y = 0.10$ (left inset), $\Delta y = 0.18$ (right inset). Black curves: dominant mean field solutions, red: MCMC calculation, blue: one-component mean field solution at $\Delta y = 0.15$.*

of the distribution. There, the MCMC calculation finds a structured behavior with three clearly distinguishable modes whereas mean field predicts a simpler pattern.

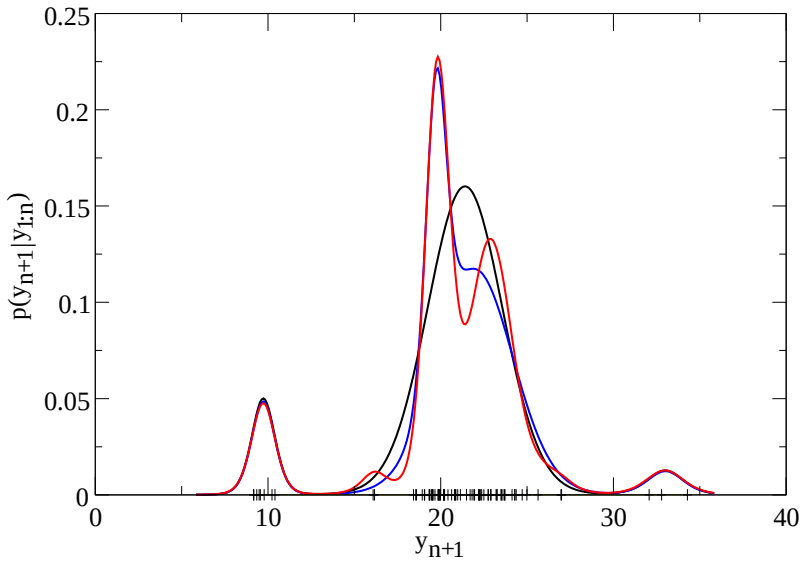

FIG 3. *Predictive density calculated from galaxy red shift data set (Roeder 1990). Parameters for the DPM were chosen as* $\alpha = 1$, $\sigma_{\text{eff}} = 0.707$, $\lambda_{\text{eff}} = 7.07$, *and* $s = 4$. *Black and blue curves show the two optimal mean field solutions which have three or four mixture components, respectively; red: MCMC calculation of DPM posterior.*

Another comparison between mean field and MCMC results is provided in Fig. 5 of Sec. 5.1. We have also studied multivariate examples using a straightforward generalization of normal/inverse-gamma DPM model described above with a diagonal covariance matrix. We have found the mean field results to show the same qualitative behavior as in the univariate case, e.g., there is still a clear distinction between occupied and unoccupied mixture components, individual observations remain firmly attached to a small number of mixture components, and the influence of the scaling parameter α is very small.

Large-n behavior of mean-field predictive density. Figure 4 provides an illustration of the behavior of the mean-field predictive density in the limit of large sample sizes. 200 random samples each containing 300 [Fig. 4(a)] and 3000 (b) observations, respectively, were drawn from the three-components normal mixture $0.3 * \mathcal{N}(0, 0.15^2) + 0.4 * \mathcal{N}(0.05, 0.04^2) + 0.3 * \mathcal{N}(-0.035, 0.02^2)$. For each sample, the predictive density was calculated based on the mean-field posterior for a location-scale DPM model with parameters $\sigma_{\text{eff}} = 0.06$, $\lambda_{\text{eff}} = 1$, $s = 0.5$, and $\alpha = 5$. To summarize the results of these calculations, Fig. 4 shows the average over the predictive densities (red curve), together with the first and third quartiles of the distribution of predicted densities at each point (green and blue curves).

The results of this simulation indicate that for growing sample size the mean-field predictive density indeed approaches the true underlying density (as long as the latter can appropriately be expressed as a normal mixture). Note that

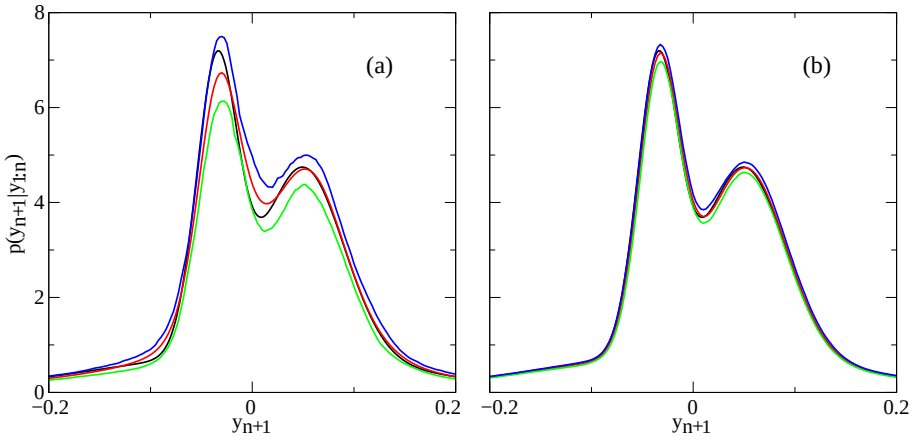

FIG 4. *Large-n behavior of mean-field predictive density.* 200 *random samples each of size* $n = 300$ *(a) and* 3000 *(b), respectively, were drawn from a three-component normal mixture with density shown by the black curve, and for each sample the mean-field predictive density was calculated. Red curves depict averaged predictive densities, whereas green and blue curves show first and third quartiles of the distribution of predicted densities at each point.*

the sample sizes necessary for obtaining a good approximation strongly depend on the details of the generating mixture, e.g., strongly overlapping normals with similar variance require many more observations than well separated mixture components.

How can the convergence behavior of the mean-field predictive density be understood? From (13), (14), and (24) follows that in the limit of $n \to \infty$, the predictive distribution in the normal location model is determined by the Bayes estimates μ_k and $\mathbb{E}(w_k)$ for the means and the weights of the occupied mixture components, only. If these quantities are estimated correctly, the predictive and the underlying true distribution will coincide since the contribution (24) of the empty components as well as the parameters ρ_k go to zero. For the normal location-scale model, the Bayes estimates for the precisions (i.e., the inverse variances) have to be taken into account as well.

An argument supporting the convergence of the predictive towards the true distribution can be given as follows. First of all, the mean field approximation appears to be able to determine the number of mixture components correctly (in the example of Fig. 4, the optimal mean field approximation had two components for about 15% of the samples with size 300, whereas for the larger sample size, mean field found three components for all samples). Furthermore, a mean field approximation with K occupied components obtained in the DPM model is very similar to the mean field result for a normal mixture with a fixed number K of components. For such a model, however, Wang and Titterington (2006) have shown that the Bayes estimates converge against the maximum likelihood estimates and hence against the true mixture parameters. This explains the good convergence behavior of the mean-field predictive density in the DPM model.

5. Alternative mean-field approximation schemes

The mean-field approximation scheme discussed in Sec. 2.2 which is based on the stick-breaking representation of the DPM model was introduced by Blei and Jordan (2006). Several variants of this method were subsequently proposed by Kurihara, Welling, and Teh (2007). In this section, we will examine one of these schemes in more detail, i.e., the substitution of the truncated stick-breaking prior by a finite-dimensional Dirichlet prior. We also briefly discuss marginalization over the weight variables v_k in the stick-breaking prior and the data likelihood parameters.

5.1. Finite-dimensional Dirichlet prior

As discussed, e.g., in Ishwaran and James (2001), replacing the truncated stick-breaking prior by a finite-dimensional Dirichlet distribution also provides an excellent approximation to the full DPM model. More specifically, in this approach one uses the probability model

$$p(y_{1:n}, z_{1:n}, w_{1:K}, \eta_{1:K}) = \prod_{i=1}^{n}\prod_{k=1}^{K}[p(y_i|\eta_k)w_k]^{\mathbb{I}(z_i=k)}p(w_{1:K})\prod_{k=1}^{K}G_0(\eta_k), \quad (29)$$

where the variables $w_{1:K}$ have a Dirichlet distribution $\mathrm{Dir}(\alpha/K, \ldots, \alpha/K)$. The cutoff K has again to be chosen large enough to make the approximation accurate.

In the mean field treatment of this model, we seek the best approximation within the class of distributions $q(z_{1:n})q(\eta_{1:K}, w_{1:K})$. The update equations for the iteration scheme imply that the optimal functions factorize further and are of the form $q(z_{1:n})q(\eta_{1:K}, w_{1:K}) = \prod_{i=1}^{n} q(z_i)\prod_{k=1}^{K} q(\eta_k)q(w_{1:K})$. Explicitly, the iteration steps are carried out as follows:

$$q(z_i = k) = \pi_{ik} \propto \exp\left\{\mathbb{E}_{\eta_k}[\log p(y_i|\eta_k)] + \mathbb{E}_{w_{1:K}}(\log w_k)\right\}, \quad (30)$$

$$q(w_{1:K}) = \mathrm{Dir}(a_1, \ldots, a_K) \text{ with } a_k = \frac{\alpha}{K} + n_k, \quad (31)$$

$$q(\eta_k) \propto \exp\left[\sum_{i=1}^{n}\pi_{ik}\log p(y_i|\eta_k) + \log G_0(\eta_k)\right]. \quad (32)$$

The expression for the lower bound is given by

$$\begin{aligned} H = {} & \sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}\mathbb{E}_{\eta_k}(\log p(y_i|\eta_k)) + \sum_{k=1}^{K}\mathbb{E}_{\eta_k}(\log G_0(\eta_k)) - \sum_{i=1}^{n}\sum_{k=1}^{K}\pi_{ik}\log\pi_{ik} \\ & - \sum_{k=1}^{K}\mathbb{E}_{\eta_k}(\log q(\eta_k)) + \log\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} + \sum_{k=1}^{K}\log\frac{\Gamma(\alpha/K+n_k)}{\Gamma(\alpha/K)}. \end{aligned} \quad (33)$$

This relation holds at convergence or at any stage of the iteration process if $q(w_{1:K})$ has been updated following an update of $q(z_{1:n})$. The calculation of the predictive density is outlined below.

TABLE 1
Mean field parameters of populated mixture components for the example of Fig. 5.

	Mean μ_k	Standard deviation ρ_k	Number of observations n_k	Weight w_k
Truncated stick-breaking				
1	−0.0282297	0.0128866	42.9635	0.348917
2	0.0251737	0.0118795	40.2126	0.323145
3	−0.00454731	0.118282	36.8239	0.289649
Lower bound $H = 138.633$				
Finite-dimensional Dirichlet				
1	−0.0281957	0.0128127	42.5747	0.342598
2	0.0251655	0.0118781	39.9802	0.321842
3	−0.00463867	0.117388	37.4451	0.301561
Lower bound $H = 134.616$ ($K = 20$)				
Truncated stick-breaking after marginalization over v_k (see Sec. 5.2)				
1	−0.0282303	0.0129080	42.8283	0.347844
2	0.0251750	0.0118822	40.0631	0.321977
3	−0.00451457	0.117859	37.1086	0.291889
Lower bound $H = 138.895$				

In order to compare this approach to the method using the truncated stick-breaking prior, we now discuss a representative numerical example that illustrates the main aspects. We consider a situation where data are drawn from two adjacent normal distributions and a superimposed broader distribution. In this case, some observations are found to be partially assigned to three different mixture components. This somewhat involved setup has been chosen since one would expect that potential differences between the methods should show up more readily than, e.g., in a situation where all cluster assignments are clear-cut.

The data for the example were generated by drawing 40 observations each from the three distributions $\mathcal{N}(\pm 0.03, 0.01^2)$ and $\mathcal{N}(0.0, 0.12^2)$, and the parameters of the DPM model were selected as $\alpha = 5$, $\sigma_{\text{eff}} = 0.06$, $\lambda_{\text{eff}} = 1$, and $s = 0.5$. Figure 5 shows the resulting predictive density distributions together with the MCMC result, as well the assignment probabilities π_{ik} for the three mixture components that are found to be populated in the optimal mean field solutions. We note that mean field recovers the three mixture components of the original distribution and, overall, agrees well with the Monte Carlo calculation. More importantly, however, the two mean field results are found to be almost indistinguishable on the scale of Fig. 5. To further illustrate the similarity, Table 1 compares the parameters of the three occupied mixture components. In all cases, the mixture components $\mathbb{E}_{\eta_k}(p(y_{n+1}|\eta_k))$ in the predictive distributions (13) and (37) are found to be practically of Gaussian shape. Table 1 shows the corresponding normal means μ_k and standard deviations ρ_k, the numbers $n_k = \sum_i \pi_{ik}$ of observations assigned to a component, the weights w_k of the mixture components, and the lower bound H.

The close correspondence between the solutions is not a particular feature of this example, but has been observed in a similar way in many other cases that were studied numerically, using univariate as well as multivariate data. In the discussion of this observation, the following points should be emphasized.

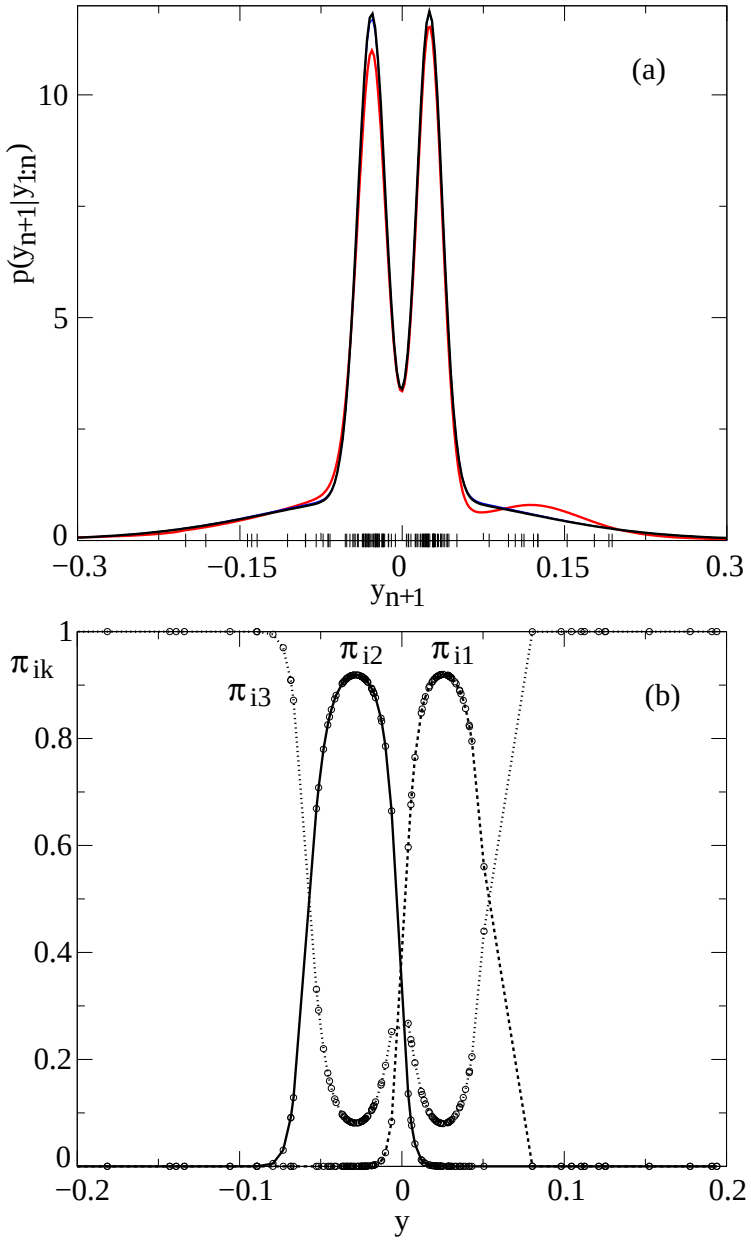

FIG 5. *Mean field approximations for three-component mixture. (a) Predictive density calculated from truncated stick-breaking priors without and with marginalization over v_k (see Sec. 5.2) (black and blue curves) and finite-dimensional Dirichlet prior (green). Red curve: MCMC calculation. Mean field results are almost indistinguishable. Parameters are $\alpha = 5$, $\sigma_{\rm eff} = 0.06$, $\lambda_{\rm eff} = 1$, and $s = 0.5$. (b) Assignment probabilities π_{ik} for occupied mixture components. Mean field results from the different schemes are indistinguishable on this scale.*

(i) The correspondence between solutions is not restricted to the global optimum, but also applies to other local maxima in H that appear as fixed points of the mean field iteration scheme. The numerical deviations between parameters in corresponding solutions, as displayed, e.g., in Table 1, are typically very small compared to the differences to other, non-corresponding solutions. In this way, correspondence can easily be established.

(ii) The reason for the close similarities between the methods can be found in the fact that they only differ in the way the prior of the DPM model is handled. The discussion of Secs. 3 and 4 has shown that the mean field solutions are mostly determined by the data-likelihood part of the DPM model whereas the influence of the prior is rather weak. The resulting similarities are thus not surprising (see also remarks in Sec. 5.2).

In spite of the overall similarity, it is nevertheless instructive to compare in more detail the use of Dirichlet and stick-breaking priors, respectively, in the mean field calculations. In order to motivate some of the general observations discussed below, we first give the result of a calculation along the lines of Sec. 3.1 for the assignment probabilities in the case of a single data point. Setting $y_1 = 0$ for convenience, one again finds mean field solutions that assign the data point to a single mixture component, i.e., one obtains assignment probabilities $\pi_{1k}^{(l)} \approx \mathbb{I}(k = l)$, $k, l = 1, \ldots, K$, or, more precisely,

$$\pi_{1l}^{(l)} \approx 1 - (K-1)\exp\left[-\frac{\lambda^2}{2\sigma^2} + \frac{1}{2(1+\frac{\sigma^2}{\lambda^2})} - \frac{K}{\alpha}\right], \tag{34}$$

$$\pi_{1k}^{(l)} \approx \exp\left[-\frac{\lambda^2}{2\sigma^2} + \frac{1}{2(1+\frac{\sigma^2}{\lambda^2})} - \frac{K}{\alpha}\right], \quad k \neq l, \tag{35}$$

where the neglected terms are at least of order ξ^2 with $\xi = \exp(-\lambda^2/2\sigma^2)$. The complete mean field solutions are obtained from these expressions using relations (31)–(32).

We now summarize some general features of mean field approximations with a finite dimensional Dirichlet prior.

(i) We again find a clear distinction between occupied and empty mixture components. In contrast to the stick-breaking case, however, the population of empty components does not remain finite (i.e., of order ξ) as $K \to \infty$, but vanishes exponentially with K (see (35)). On the other hand, if there is more than one occupied component, the corresponding assignment probabilities appear to converge much more slowly towards their asymptotic limits (with $1/K$). However, their overall variation with K is still rather small.

(ii) Permuting the component indices in a mean field solution leads to another valid solution (i.e., fixed point) of the iteration scheme (see (34)–(35)). This is a consequence of the symmetry of the Dirichlet prior. All these solutions have the same lower bound H. With the stick-breaking prior, the largest bound is obtained if the populations of components are size-ordered, i.e., $n_1 \geq n_2 \geq \ldots$ (see (12)).

(iii) With a finite-dimensional Dirichlet prior, further types of mean field solutions can arise that have no equivalents in the stick-breaking approach. For example, the single observation considered above can be assigned to more than one component. If k_0 components become occupied, then one has $\pi_{1k} \approx 1/k_0$ for these components. In the extreme case of $k_0 = K$, $\pi_{1k} = 1/K$ is an exact solution. Nevertheless, in the numerical calculations for more complex problems we have only rarely encountered solutions that were significant in terms of their lower bound and did not seem to have a correspondence in the stick-breaking approach.

(iv) The predictive density is given by

$$p(y_{n+1}|y_{1:n}) = \sum_{k=1}^{K} \mathbb{E}_{w_{1:K}}(w_k)\mathbb{E}_{\eta_k}(p(y_{n+1}|\eta_k)). \tag{36}$$

Since $\mathbb{E}_{w_{1:K}}(w_k) = (n_k + \alpha/K)/(\alpha + n)$, the evaluation of the predictive density can show an appreciable dependence on K, in particular for smaller values of K (and large α). To eliminate this dependence, it is convenient to take the asymptotic limit

$$p(y_{n+1}|y_{1:n}) \overset{K\to\infty}{\longrightarrow} \sum_{k=1}^{K} \frac{n_k}{\alpha + n}\mathbb{E}_{\eta_k}(p(y_{n+1}|\eta_k)) + \frac{\alpha}{\alpha + n}p_{\text{new}}(y_{n+1}) \tag{37}$$

Similar to the discussion of Sec. 3.2, the term $\alpha(\alpha + n)^{-1}p_{\text{new}}(y_{n+1})$ is due to the empty components and describes the assignment of y_{n+1} to a cluster of its own. It agrees precisely with the corresponding term in the exact treatment. For the practical evaluation of (37), we can make use of the fact that even for rather moderate values of K, the population of empty components is exponentially small, whereas the population of occupied components depends only weakly on K. It is therefore possible to accurately calculate (37) with a small value of K already. As an example, the predictive density in Fig. 5(a) was obtained in this way and is almost indistinguishable from the stick-breaking result.

(v) A rather interesting observation concerns the behavior of the lower bound H, which does not become asymptotically constant as in the stick-breaking case. To study this aspect in more detail, let us consider a mean field solution for which the first k_0 components are occupied and all others empty. In this case, all terms in (33) will become constant as $K \to \infty$, besides the first k_0 members of the last sum for which we find that

$$\begin{aligned} &\sum_{k=1}^{k_0}\left[\log\Gamma\left(\frac{\alpha}{K} + n_k\right) - \log\Gamma\left(\frac{\alpha}{K}\right)\right] \overset{K\to\infty}{\longrightarrow} \sum_{k=1}^{k_0}\log\Gamma(n_k) - k_0\log\Gamma\left(\frac{\alpha}{K}\right) \\ \approx\ &\sum_{k=1}^{k_0}\log\Gamma(n_k) + k_0(\log\alpha - \log K) \end{aligned}$$

after using that $\Gamma(z) \approx 1/z$ for $z \to 0$. The lower bound thus diverges as $-k_0 \ln K$. The reason for this behavior is the permutation symmetry of the

Dirichlet prior. As mentioned in (ii), this symmetry gives rise to the existence of multiple equivalent mean field solutions that differ from each other only in the permutation of component indices, and the proper mean field description should take all of them into account.

For a special case, we can explicitly show how the combination of all these solutions re-establishes the constancy of the asymptotic bound. Assume again that there are k_0 occupied mixture components. For large K, the population of the empty components will be vanishingly small as mentioned in (ii). There are thus $K!/(K-k_0)!$ index permutations leading to *distinct* mean field solutions. Now assume in addition that for each occupied mixture k there is at least one observation i assigned (almost) exclusively to this component, i.e., $\pi_{ik} \approx 1$. Such types of mean field solutions are not uncommon, examples are shown in Figs. 1 and 2. Any two index permutations of such solutions fulfill, to a very good degree of approximation, the non-overlap condition of Sec. 2.3. We can therefore calculate the lower bound H_{mix} of an equi-weighted mixture of all permutations. Since it surpasses the bound of an individual solution by $\ln K!(K-k_0)!$, H_{mix} will be asymptotically constant. We note that the approach to the limit is of order $1/K$ and thus much slower than in the stick-breaking case. A very similar behavior is observed for the exact calculation of the lower bound in the truncated DPM model with Dirichlet prior.

In the general case, however, the mean field solution will not fulfill the non-overlap condition (see, e.g., Fig. 5). It is then no longer possible to calculate the entropy term of the lower bound analytically. It is thus an interesting and challenging problem to determine the asymptotic behavior of the lower bound under these circumstances.

The divergent behavior of the lower bound can be of relevance for practical calculations. The scaling with $-k_0 \ln K$ implies that for growing K mean field solutions with small k_0 will appear more and more favorable. In fact, in many cases there exists a solution that assigns all observation to the same mixture component (i.e., $k_0 = 1$). Such a solution will always dominate in the limit $K \to \infty$. It is thus essential to be aware of this divergent behavior when ranking mean field solution based on their lower bound. On the other hand, one should also keep in mind that apart from the problem with the lower bound, finite-dimensional Dirichlet and truncated stick-breaking priors essentially yield the same approximations to the DPM posterior (Ishwaran and Zarepour 2002). Since the latter approach does not suffer from problems regarding the lower bound and the ranking of solutions, it might be a more convenient choice in practical calculations.

Finally, we briefly mention recent work on the latent Dirichlet allocation (LDA) model (Blei, Ng, and Jordan 2003). On a formal level, LDA is a generalization of a mixture model where the data likelihood is given by a discrete multinomial distribution and thus shares some similarity with the models described in this section. Since mean field methods are also often applied to the LDA, we expect that some of the results discussed here, e.g., regarding the structure of the solutions and the effects of the permutation symmetry of the Dirichlet prior, might be of relevance in this context as well.

5.2. Other approximation schemes

Marginalization over weight variables. As discussed by Kurihara, Welling, and Teh (2007), one might hope to improve the mean-field approximation by integrating out the v_k's of the stick-breaking representation so that the contribution of the prior weights is treated exactly. As a representative illustration of the results that are obtained from this method, we again consider the example of Sec. 5.1. Figure 5 and Table 1 show the predictive density and the numerical values for the parameters of the mixture components, respectively. Again, we find a close similarity to the results from the other approaches. This correspondence between solutions has also been observed in all other numerical studies and not only applies to the globally optimal mean field result, but also to other local maxima with large H. As explained above, we attribute this behavior to the very weak influence of the prior distribution in the mean field calculations.

For corresponding solutions, the lower bound H is larger in the marginalized model, as one would expect since the prior weight variables v_k have been treated exactly (a detailed study of the improvement in H due to marginalization has been presented by Mukherjee and Blei (2009) in the context of latent Dirichlet allocation). However, given the close similarity between the actual mean field distributions, as exemplified in Table 1, the increase in H does not seem to make the marginalized method preferable, in particular since the numerical computation of the update rules becomes more expensive. This can be avoided by using approximation schemes (Kurihara, Welling, and Teh 2007), but then the lower bound is no longer guaranteed to increase in each iteration step.

Marginalization over data likelihood parameters. A further variant of the mean field method that is mentioned, although not elaborated on, in Kurihara, Welling, and Teh (2007) consists in marginalizing over the mixture component parameters η_k. However, in this case the computational cost of exactly evaluating the mean-field update relations is increased dramatically and it is not clear if efficient approximations can be derived. Nevertheless, it is still of interest to investigate the general behavior of this approach in the case of small samples and compare it to the other versions. The main result of our studies is that the mean field solutions obtained in this scheme do not have equivalents in the other approaches, but are clearly distinct. This confirms our conclusion that the similarity of the previous methods is due to the fact that they only differ in how the prior is handled, and that the prior only has a small effect on the resulting approximation. In particular, it is found that after marginalizing over η_k the observations tend to be spread out over more mixture components; however, the inherent limitations of the factorized form for $q(z_{1:n})$, as outlined in Sec. 8, prevent this approach from yielding a significantly improved approximation to the distribution of mixture components.

6. Mean field inference for the parameter α

The DPM model can be extended to include a prior distribution $p(\alpha)$ for the parameter α. In order to preserve conjugacy, we will in the following choose $p(\alpha)$ as

a gamma distribution with shape and inverse scale parameters s_1 and s_2, respectively. To derive the mean field iteration scheme for this model, we again apply a truncation as in (1). Analytical update rules are obtained if we seek the mean field approximation within the class of distributions $q(z_{1:n})q(\eta_{1:K}, v_{1:K-1})q(\alpha)$. As in Sec. 2.2, the optimized distributions factorize in all variables. More specifically, the mean field posterior for α remains a gamma distribution with update equations given by (Blei and Jordan 2006)

$$s_1^* = s_1 + K - 1, \tag{38}$$

$$s_2^* = s_2 - \sum_{k=1}^{K-1} \mathbb{E}_{v_k} \left(\ln(1 - v_k) \right), \tag{39}$$

whereas the update for the v_k's is modified to

$$v_k \sim \text{Beta} \left(\alpha_k = n_k + 1, \beta_k = n_k^> + \mathbb{E}_\alpha(\alpha) \right). \tag{40}$$

Similarly, in the update rule (8), α has to be replaced by $\mathbb{E}_\alpha(\alpha)$, whereas the relation for $q(\eta_k)$ remain unchanged. The lower bound at convergence is now given by

$$\begin{aligned} H = & \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{ik} \mathbb{E}_{\eta_k}(\log p(y_i|\eta_k)) + \sum_{k=1}^{K} \mathbb{E}_{\eta_k}(\log G_0(\eta_k)) - \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{ik} \log \pi_{ik} \\ & - \sum_{k=1}^{K} \mathbb{E}_{\eta_k}(\log q(\eta_k)) + \log \frac{\Gamma(\mathbb{E}(\alpha) + n_K)}{\Gamma(\mathbb{E}(\alpha) + n)} + \sum_{k=1}^{K-1} \log \frac{\Gamma(n_k + 1)}{n_k + n_k^> + \mathbb{E}(\alpha)} \\ & - s_1^* \log s_2^* + s_1^* + \log \Gamma(s_1^*) - s_2 \mathbb{E}(\alpha) - \log \Gamma(s_1) + s_1 \log s_2. \end{aligned} \tag{41}$$

Some numerical experiments indicate that the convergence through standard iteration for $q(\alpha)$ is quite slow and takes longer than for the other parameters.

When a prior distribution for α is included in the model, one finds that the role of the truncation level K becomes more complex than before. First of all, (38) shows that the mean field posterior for α more and more approaches a normal distribution as K and hence the shape parameter s_1^* increase, irrespective of any other aspects of the system under study. The mean of the posterior distribution is given by $\mu = \frac{s_1^*}{s_2^*}$ and its standard devation by $\rho = \sqrt{\mu / s_2^*}$.

Under the assumption of unoccupied higher-order mixture components, one can show that $\mathbb{E}(\alpha)$ is independent of K. Using the relations from the update equations, one has at convergence

$$\mathbb{E}(\alpha) = \frac{s_1^*}{s_2^*} = \frac{s_1 + K - 1}{s_2 + \sum_{k=1}^{k_0} [\psi(\alpha_k + \beta_k) - \psi(\beta_k)] + (K - 1 - k_0)/\mathbb{E}(\alpha)}. \tag{42}$$

Here, we have assumed that the first k_0 components are populated, and used the relation $\psi(x + 1) - \psi(x) = 1/x$. The parameters α_k and β_k are defined in (40). Solving for $\mathbb{E}(\alpha)$ yields

$$\mathbb{E}(\alpha) = \frac{s_1 + k_0}{s_2 + \sum_{k=1}^{k_0} [\psi(\alpha_k + \beta_k) - \psi(\beta_k)]}, \tag{43}$$

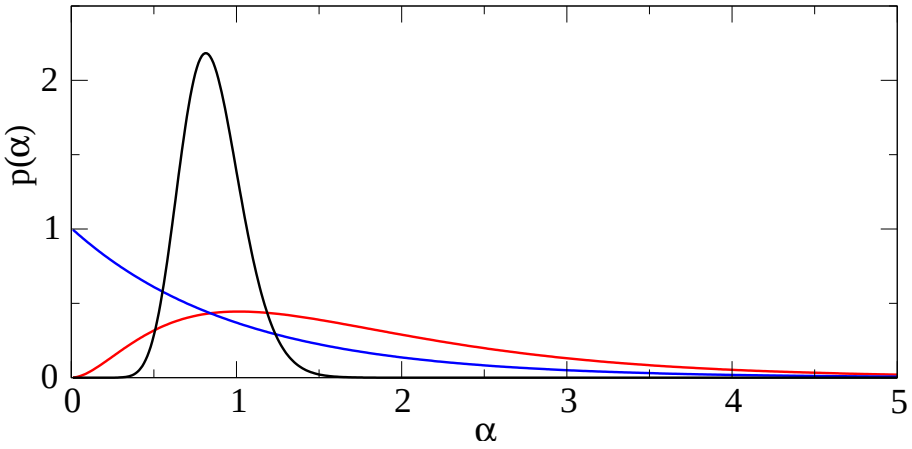

FIG 6. *Exact prior (blue curve) and posterior (red) distributions for the parameter α together with mean field approximation (black) at $K = 20$ for a sample with $n = 12$. Due to the smallness of the sample the exact posterior can be calculated without using Monte-Carlo integration. In this example, the variational posterior mean $\mathbb{E}(\alpha)$ is smaller than the prior one ($\mathbb{E}_{\text{prior}}(\alpha) = 1$), whereas the exact one is larger. For $K \to \infty$, $q(\alpha)$ approaches a Dirac delta function at $\mathbb{E}(\alpha)$.*

i.e., $\mathbb{E}(\alpha)$ is independent of K. Note that β_k depends explicitly on $\mathbb{E}(\alpha)$, and that both α_k and β_k depend implicitly on it through n_k and $n_k^>$. The latter dependence is quite weak, however. Altogether, (43) is thus an implicit equation for $\mathbb{E}(\alpha)$. It is also easy to see from (43) that the result for $\mathbb{E}(\alpha)$ is independent of the precise choice of k_0 as long as all occupied mixture components are included in the summation (i.e., one might change $k_0 \to k_0 + \Delta k$, $\Delta k > 0$). Since the iteration relations for the other factors in the mean field approximation depend on $q(\alpha)$ only through $\mathbb{E}(\alpha)$, it is ensured the assignment probabilities π_{ik} will also remain independent of K.

In the limit of $K \to \infty$, it follows from the above discussions that s_1^* and s_2^* both go to infinity, since their ratio $\mathbb{E}(\alpha)$ has to remain fixed. The approximately Gaussian posterior $q(\alpha)$ will thus be more and more localized around $\mathbb{E}(\alpha)$, i.e., it tends towards a Dirac delta function $\delta(\alpha - \mathbb{E}(\alpha))$.

We also find that the lower bound no longer attains a constant limit for $K \to \infty$. More specifically, it follows from (41) that H diverges as $-\frac{1}{2}\ln K$ in the limit of $K \to \infty$. To derive this result, one uses the constancy of $\mathbb{E}(\alpha)$ with K which allows to replace s_2^* by $s_1^*\mathbb{E}(\alpha)$, together with Stirling's approximation, and the fact that components are occupied only up to a fixed index $k_0 < K$. The divergence is caused by the entropy term $-\int q(\alpha)\ln q(\alpha)$. Qualitatively, the contracting $q(\alpha)$ more and more deviates from the exact posterior marginal and thus leads to a larger Kullback-Leibler divergence. We note that this example shows how a seemingly minor modification of the model can lead to a rather strong change in the properties of the mean field approximation, here in the form of the behavior of the lower bound. The contraction of $q(\alpha)$ and the

divergence of the lower bound can probably be avoided if one seeks the mean field approximation within the class of distributions $q(z_{1:n})q(\eta_{1:K})q(v_{1:K-1},\alpha)$. However, in this case the update rules are no longer of simple closed form but require numerical integrations.

In the numerical examples studied, the variational mean $\mathbb{E}(\alpha)$ was always smaller than the exact one. In Fig. 6 we display a case where the mean-field result not only predicts a wrong shape of the posterior, but even gives a wrong direction for the shift from prior to posterior mean (i.e., smaller instead of larger value for expectation value). One thus has to be very careful when drawing conclusions about the exact behavior based on variational results.

7. Mean field approximation and MAP estimation

An interesting perspective on the mean-field approximation is obtained from studying its connection to maximum a-posteriori (MAP) estimation of the posterior mixture distribution. We define the MAP estimation problem as finding those values of $\eta_{1:K}$ and $v_{1:K-1}$ (and hence the specific mixture distribution defined by them) that maximize the truncated posterior distribution $p(\eta_{1:K},v_{1:K-1}|y_{1:n})$ for given observations $y_{1:n}$. An algorithm for determining the MAP estimate can be constructed in close analogy to the expectation-maximization (EM) method, with the assignment variables z_i playing the role of the hidden variables in the EM algorithm. More specifically, a cycle in the EM-type iteration scheme for the normal location model is carried out by alternatingly calculating

$$\pi_{ik} \propto p(y_i|\eta_k)v_k\prod_{l=1}^{k-1}(1-v_l) = \exp\left[\log p(y_i|\eta_k) + \log v_k + \sum_{l=1}^{k-1}\log(1-v_l)\right] \quad (44)$$

for given $\eta_{1:K}$ and $v_{1:K-1}$, and updating $\eta_{1:K}$ and $v_{1:K-1}$ according to

$$v_k = \frac{\sum_{i=1}^n \pi_{ik}}{\sum_{l=k}^K\sum_{i=1}^n \pi_{il}+\alpha-1} = \frac{n_k}{n_k+n_k^>+\alpha-1}, \quad k<K, \quad (45)$$

$$\eta_k = \frac{\sum_{i=1}^n \pi_{ik}y_i}{\sigma^2/\lambda^2+n_k}. \quad (46)$$

To simplify the following discussion, we now assume $\alpha>1$. At a fixed point of the iteration scheme, we have $v_k = n_k/(n_k+n_k^>+\alpha-1)$ and

$$\pi_{ik} = \frac{p(y_i|\eta_k)[n_k+\mathbb{I}(k=K)(\alpha-1)]}{\sum_{l=1}^K p(y_i|\eta_l)[n_l+\mathbb{I}(l=K)(\alpha-1)]}. \quad (47)$$

From this, we obtain an implicit condition for the component occupation numbers n_k from the summation

$$n_k = \sum_{i=1}^n \pi_{ik} = [n_k+\mathbb{I}(k=K)(\alpha-1)]\sum_{i=1}^n \frac{p(y_i|\eta_k)}{\sum_{l=1}^K p(y_i|\eta_l)[n_l+\mathbb{I}(l=K)(\alpha-1)]}. \quad (48)$$

The solutions of the MAP iteration scheme have the following general properties:

1. Given a solution, we can construct $(K-1)!-1$ further solutions by permuting the indices of the components with $k \neq K$. To see this, consider such a permutation $\mathcal{P}$ that keeps K invariant, i.e., $\mathcal{P}(K) = K$. From the given fixed-point solution, we can construct the new solution starting from $\pi_{ik}^{(\mathrm{new})} = \pi_{i\mathcal{P}^{-1}(k)}^{(\mathrm{old})}$ so that $n_k^{(\mathrm{new})} = n_{\mathcal{P}^{-1}(k)}^{(\mathrm{old})}$. This implies that $\eta_k^{(\mathrm{new})} = \eta_{\mathcal{P}^{-1}(k)}^{(\mathrm{old})}$ from (46). We now choose $v_k^{(\mathrm{new})} = n_k^{(\mathrm{new})}/(n_k^{(\mathrm{new})} + n_k^{>,\,(\mathrm{new})} + \alpha - 1)$. Re-evaluating π_{ik} using (44) proves the consistency of the new solution. However, if the index K were included in the permutation as well, the consistency would no longer hold. Note that all possible $(K-1)!$ permutations describe the same mixture distribution defined by the (unordered) set of η_k's and their associated weights $w_k = v_k \prod_{j=1}^{k-1}(1-v_j)$.

2. All $(K-1)!$ solutions obtained above have the same posterior probability, which is immediately seen from the relation

$$\sum_{z_{1:n}} p(y_{1:n}, z_{1:n}, v_{1:K-1}, \eta_{1:K}) = \prod_{i=1}^{n}\left(\sum_{k=1}^{K} p(y_i|\eta_k) w_k\right) \prod_{k=1}^{K-1} \alpha^{K-1} w_K^{\alpha-1} \prod_{k=1}^{K} G_0(\eta_k). \tag{49}$$

Expression (49) also explains the special role of the component K mentioned above. The mixture component with the largest weight has to be assigned to label K in order to maximize the posterior probability (note that we have assumed $\alpha > 1$). The assignment of the other components is arbitrary. For all the other $K! - (K-1)!$ ways in which the mixture distribution can be represented with w_K not being the maximum weight, the posterior probability is smaller. Numerical calculations indicate that these representations are not even local maxima of the posterior.

3. The MAP equations may have multiple fixed-point solutions corresponding to different mixture distributions. However, for any set of observations $y_{1:n}$, $\pi_{ik} = \mathbb{I}(k = K)$ (together with the ensuing relations for $v_{1:K-1}$ and $\eta_{1:K}$) is always a fixed point of the MAP iteration scheme. In this case, all observations are assigned to the mixture component K. In agreement with the above discussion, $\pi_{ik} = \mathbb{I}(k = l)$, $l < K$, can never be a fixed point as one easily sees from the iteration equations or from the fact that the posterior probability vanishes in this case.

4. As in mean field, the MAP iteration scheme supports solutions for which only a subset of the mixture components are occupied. The populations of the empty modes vanish exactly (i.e., $\pi_{ik} = 0$), whereas in mean field they are only exponentially suppressed. The exact vanishing of component populations can be understood from two perspectives. On the one hand, given a MAP solution, we can immediately construct an equivalent solution with a larger number of components simply by inserting components with zero weight. Second, the implicit relation (48) for n_k has the trivial solution $n_k = 0$, $k < K$, and in many cases this solution is indeed assumed.

There are several strong indications for a close connection between the mean field method and MAP estimation. First of all, the latter can also be derived from a mean-field-type approximation to the posterior DPM distribution. To this end, one simply restricts the trial functions to the set (MacKay 2003, Ch. 33)

$$q(z_{1:n}, \tilde{v}_{1:K-1}, \tilde{\eta}_{1:K}) = q(z_{1:n}) \prod_{k=1}^{K-1} \delta(\tilde{v}_k - v_k) \prod_{k=1}^{K} \delta(\tilde{\eta}_k - \eta_k). \tag{50}$$

The mean field update equations for $q(z_{1:n})$ and the location parameters v_k, η_k of the Dirac delta functions are identical to the EM equations derived above. A lower bound can be calculated in the usual way after setting the entropy associated with the delta-distributions equal to zero. In this sense, MAP estimation can be considered equivalent to a restricted mean field approximation problem. In (50), the distributions of $\tilde{v}_{1:K-1}$ and $\tilde{\eta}_{1:K}$ are completely localized. In the full mean field scheme, populated components similarly tend to become localized (see expressions (9) and (11)). We also note the correspondence between the actual update equations: expression (11) for μ_k is identical to (46), the value of v_k in (45) matches the mode of the beta distribution (9), and the iterations for $q(z_{1:n})$ given by (7) and (44) are also very similar to each other.

Second, numerical studies often show strong similarities between MAP and mean field solutions. For example, the assignment probabilities π_{ik} calculated in the respective iteration schemes are typically found to be very close to each other. Third, in a closely connected problem, Wang and Titterington (2006) have shown that for a normal mixture model with a fixed number of components, the variational Bayesian point estimates asymptotically converge against the maximum likelihood estimates.

The above arguments show that the mean field approximation to the DPM model is closely related to the MAP estimation of the mixture distribution. For a qualitative picture, we can envisage the space of all mixture distributions defined by the (unordered) set of mixture weights w_k and parameters η_k. MAP estimation singles out a specific mixture, whereas mean field implies a density distribution in this space that is smeared out around this selected mixture. This connection also gives a qualitative explanation of why the mean field approximation is compressed into a small number of mixture components, as it inherits this property from the MAP solution.

Often, however, MAP estimates provide a poor representation of the behavior of the full DPM probability distribution, in particular with regard to clustering. For example, for a set of localized observations $|y_i| \lesssim \sigma$, the MAP estimate is often given by the single-component solution $\pi_{ik} = \mathbb{I}(k = K)$. It thus fails completely to describe the posterior distribution of the number of mixture components. It is therefore not surprising that the mean field approximation suffers from the same problem.

To get some insight into the reason for this failure, it is instructive to consider how the component number distribution in the DPM model comes about. It can be obtained, at least in principle, from (2) which decomposes the marginal $p(y_{1:n})$ into a sum of the contributions from all possible partitionings of the data

into clusters. In the case of localized observations $|y_i| \lesssim \sigma$ and moderate α, it is often found that the term that assigns all observations to a single cluster has the largest probability whereas the individual probabilities for all other clusterings are significantly smaller. However, since there is a large number of possibilities for putting the observations into two or more clusters, the resulting distribution for the number of components can still become quite broad. Although not strictly in one-to-one correspondence, the fact that the single-cluster assignment has the largest individual probability indicates that the MAP estimate should also put all observations into a single mixture component.

Finally, we address the fact that the MAP and mean field solutions differ distinctively in how indices are assigned to the populated mixture components. As discussed above, in MAP the component with largest weight has index K. In the mean field approximation the lower bound is maximized if the populations of components are ordered by size, i.e., $n_1 > n_2 > \dots$ (see (12)).

We explain this difference in behavior as follows. In the stick-breaking approach, a mixture distribution can be represented in $K!$ different ways by appropriately choosing v_k's and η_k's. With MAP, we select a representation that maximizes posterior probability density. The mean field solution can be interpreted as providing a probability distribution in mixture space which is localized around the *same actual mixture* as chosen by MAP. However, it puts its weight on a *different representation* of the mixture, namely one with a large associated probability mass (with respect to a volume element in mixture space). Such a representation typically assigns small indices k to all components with large mixture weights.

8. Summary and conclusions

Recently, variational algorithms, and in particular mean field methods, have received increasing attention as possible alternatives to MCMC integration in computational Bayesian inference. In this paper, we have presented a systematic investigation of several mean field inference schemes for Dirichlet process mixture models that were proposed in Blei and Jordan (2006) and Kurihara, Welling, and Teh (2007). Our discussion focussed on the normal location and normal/inverse-gamma location-scale models which we believe to be among the most important for practical applications of mean field techniques in this context. The main results of our studies can then be summarized as follows.

(i) In typical fixed point solutions to the mean field iteration scheme, there is a clear distinction between occupied and essentially empty mixture components. The population of empty components is exponentially suppressed. Individual observations tend to be firmly attached to a relatively small number of components, often even only a single one. This behavior was explained in terms of a "self-reinforcement" effect in the iteration equations. We have discussed several structural properties of the solutions. To illustrate this discussion, explicit approximations for the case of a single observation were presented. It was also studied how the mean field approximation chooses between different clusterings

of observations. As another important feature of the mean field solutions, their very weak dependence on the DPM parameter α was pointed out.

(ii) We compared different variants of the mean field scheme that are derived from the unmarginalized and marginalized truncated stick-breaking priors and the finite-dimensional Dirichlet priors, respectively. It was found that these methods lead to essentially equivalent results regarding density estimation and cluster allocation. This virtual equivalence was attributed to the very weak influence of the prior distribution in the mean field iteration scheme. From a practical point of view, the unmarginalized stick-breaking variant appears to be most convenient for numerical work. In the marginalized version, one either is faced with increased computational cost in the mean field iteration scheme or has to make use of approximation methods. For the finite-dimensional Dirichlet prior, we have pointed out the divergent asymptotic behavior of the lower bound and related this effect to the permutation symmetry of the Dirichlet prior. The divergence may cause problems when ranking mean field solutions based on their lower bound.

(iii) We have clarified the role of the truncation level K. The mean field solutions display a well-defined behavior in the limit of $K \to \infty$. Together with the fact that only a fixed number of mixture components becomes occupied, this allows to characterize the asymptotic behavior of important quantities such as the lower bound and the predictive density. For practical calculations, it is sufficient to choose K slightly larger than the number of occupied components to calculate the asymptotic quantities. It thus seems natural to work with the asymptotic limit of the mean field approximation, rather than treating K as an additional variational parameter. This approach also keeps in line with the fact that the full DPM model arises as the asymptotic limit of the truncated model.

(iv) When a prior distribution for α is included in the model, the role of the truncation level becomes more complex. Whereas $\mathbb{E}(\alpha)$ becomes constant in the limit of $K \to \infty$, the mean field posterior for α approaches a Dirac delta function, and the lower bound H decreases without bound. This behavior appears to be an artifact of the factorization assumptions in the mean field trial functions and, on a qualitative level, is not easily reconciled with the fact that the mean field solutions without the α prior only depend very weakly this parameter.

(v) It was shown that the mean field approximation is closely related to MAP estimation of the DPM model. The MAP estimation problem can be shown to be equivalent to mean field inference under a restricted set of trial functions. The unrestricted mean field approximation can be thought of as providing a distribution over mixture space that is smeared out around the mixture singled out by the MAP estimate.

(vi) Compared with MCMC calculations, mean field results for the predictive density distribution were often found to be quite accurate. An explanation of this behavior in the limit of large sample sizes was given at the end of Sec. 4. However, there can be strong discrepancies regarding clustering and the number of posterior mixture components.

A well-known shortcoming of the mean-field method concerns the fact that it typically underestimates the variance of the target distribution (see, e.g., Bishop (2006), Sec. 10.1; MacKay (2003), Ch. 33). This limitation is due to the fact that the factorized form of the approximation cannot properly capture the correlation structure present in the full model. This problem is also present in the context of the DPM model and accounts for some of the behavior discussed in the previous sections.

As a simple illustration, consider the marginal prior distribution $p(z_1, z_2)$ in the finite-dimensional Dirichlet model of Sec. 5.1 with $n = 2$ observations. The distribution takes only two values, depending on whether $z_1 = z_2$ or not, but this behavior cannot be modeled by a factorization $q(z_1)q(z_2)$. For a larger number of observations, the correlations become even more complicated. A further example is provided by the case of two well-separated clusters of observations. In this case, there is strong anticorrelation in the cluster assignment, in that observations from both clusters can be assigned to the mixture component with label k, but not at the same time (i.e., the posterior marginals $p(z_i = k)$ and $p(z_j = k)$ can both be large, but $p(z_i = z_j = k)$ vanishes if i and j belong to different clusters). In factorized mean field, the observations from the two clusters always have to be assigned to different mixture components, $q(z_i = k)$ and $q(z_j = k)$ cannot be large simultaneously.

In this context, it is interesting to note that for the stick-breaking approach of Sec. 2.2, the factorized form is optimal even within the large class of trial functions $q(z_{1:n})q(v_{1:K-1}, \eta_{1:K})$ (and similarly for the finite-dimensional Dirichlet prior). An important task of future work thus consists in finding more general, tractable classes of trial functions that permit improvements to the mean field scheme. In the present study, we have already considered mixtures consisting of different fixed point solutions to the mean field iteration scheme. These mixtures were mathematically tractable because of the particular structure of the fixed point solutions. However, this approach did not lead to a substantial improvement in the quality of the approximation.

As an overall conclusion, we can state that the mean field method, when viewed on its own, produces reasonable results regarding density estimation and clustering. This can be explained by its connection to MAP estimation of mixture distributions. In view of point (vi) above, however, care is needed when it is used as an actual approximation to the exact DPM posterior. Nevertheless, since mean field appears to be a very useful and efficient way of calculating inferences for large-scale DPM problems it is very important to have a solid understanding of its properties. It is hoped that the present paper makes some contributions in this direction. We expect that some of the results may be also of interest in related contexts, such as latent Dirichlet allocation which is often treated with mean field techniques as well Blei, Ng, and Jordan (2003).

Acknowledgements

This work was supported by an EPSRC Statistics Mobility grant. The author would like to thank Prof. P. Green for a careful reading of the manuscript.

Helpful comments and advice by the associate editor and an anonymous referee are also gratefully acknowledged.

Appendix: Characteristic features of mean field solutions

In this Appendix, we give some brief arguments in order to explain the characteristic features of the mean field solutions outlined in Sec. 3.2. In the limit of $\lambda \to \infty$ at fixed σ, i.e., $\sigma/\lambda \to 0$, the fixed points of the mean field iteration scheme are determined by

$$\pi_{ik} \propto \exp\left[-\frac{1}{2n_k} - \frac{(y_i - \mu_k)^2}{2\sigma^2} + \psi(n_k + 1) - \sum_{j=1}^{k-1} \frac{1}{n_j^> + \alpha}\right] \tag{51}$$

with $\mu_k = \sum_i \pi_{ik} y_i / n_k$. The appearance of the term $-1/2n_k$ in the exponent of (51) shows that the iteration scheme can entertain solutions with exactly vanishing population in some components, i.e., $n_k = 0$. However, such solutions will be nonanalytic in σ/λ when we consider their variation with this parameter. Given any solution of the iteration scheme (51), consider its "pruned" version where all empty components have been discarded. The pruned version will be the solution of an iteration scheme in which only the occupied components are retained. Solutions of the pruned scheme in fact exist; e.g., for $K = 1$ solutions are given by $\pi_{i1} = 1$, and for $n = 2$, $K = 2$ they can be found graphically. By construction, all pruned π_{ik} will be non-vanishing. If we re-establish the λ dependence of the pruned scheme, we can expect the assignment probabilities to smoothly depend on σ^2/λ^2 by the theorem of implicit functions.

The effect of adding the empty components back into the iteration scheme at finite σ/λ can be seen from a perturbative argument. Calculating the assignment probabilities for the unoccupied components from the pruned π_{ik} in a first iteration step shows that the former will be of order $\exp(-\lambda^2/2\sigma^2)$, i.e., exponentially small compared to the pruned π_{ik} for sufficiently small σ^2/λ^2. Further iteration steps will only provide negligible corrections to the result of the first iteration. The scaling with $\exp(-\lambda^2/2\sigma^2)$ reflects the nonanalytic behavior of the unoccupied assignment probabilities. From the term $-\sum_{j=1}^{k-1}(\alpha + n_j^>)^{-1}$ in the exponent of (8) follows that the π_{ik}'s for the unoccupied components scale with $\exp(-k/\alpha)$ as soon as their index k is larger than the largest index for occupied components. The geometric scaling with k implies that the combined weight of the unoccupied components remains bounded as $K \to \infty$ which leads to the insensitivity of all assignment probabilities with the truncation level.

References

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174. MR0365969

Bishop, C.M. (2006). *Pattern recognition and machine learning.* Springer, New York. MR2247587

Blei, D.M. and Jordan, M.I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1**, 121–143. MR2227367

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet allocation. *J. Mach. Learning Res.* **3**, 993–1022.

Escobar, M.D. (1988). Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Yale University, Department of Statistics.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89**, 268–277. MR1266299

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577–588. MR1340510

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230. MR0350949

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96**, 161–173. MR1952729

Ishwaran, H. and Zarepour, M. (2002). Exact and approximate sum representations for the Dirichlet process. *Canad. J. Statist.* **30**, 269–283. MR1926065

Jaakkola, T.S. and Jordan, M.I. (1998). Improving the mean field approximation via the use of mixture distributions. In *Learning in Graphical Models*, ed. M.I. Jordan, MIT Press, Cambridge, MA, 163–174.

Kurihara, K., Welling, M., and Teh, Y.W. (2007). Collapsed variational Dirichlet process mixture models. In *Proceedings of IJCAI-07*, 2796–2801.

Kurihara, K., Welling, M., and Vlassis, N. (2007). Accelerated variational Dirichlet process mixtures. In *Advances in Neural Information Processing Systems, Vol.* **19**, eds. B. Schölkopf, J.C. Platt and T. Hofmann, MIT Press, Cambridge, MA, 761–768. MR2441316

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12**, 351–357. MR0733519

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Comm. Statist. Simulation Comput.* **23**, 727–741. MR1293996

MacEachern, S.N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7**, 223–238.

MacKay, D.J.C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press, New York. MR2012999

Mézard, M., Parisi, G., and Virasoro, M.A. (1987). *Spin glass theory and beyond*. Lecture Notes in Physics, Vol. **9**. World Scientific Publishing, Teaneck, NJ. MR1026102

Mukherjee, I. and Blei, D.M.(2009). Relative performance guarantees for approximate inference in latent Dirichlet allocation. In *Advances in Neural Information Processing Systems* **21**, ed. D. Koller, Y. Bengio, D. Schuurmans, L. Bouttou, and A. Culotta, 1129–1136.

Neal, R.M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9**, 249–265. MR1823804

Opper, M. and Saad, D.(eds.) (2001). *Advanced mean field methods: theory and practice*. Neural Information Processing Series. MIT Press, Cambridge, MA. MR1863214

Roeder, K.(1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* **85**, 617–624.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4**, 2, 639–650. MR1309433

Teh, Y.W., Kurihara, K., and Welling, M. (2008). Collapsed variational inference for HDP. In *Advances in Neural Information Processing Systems* **20**, ed. J.C. Platt, D. Koller, Y. Singer, and S. Roweis, Cambridge, MA: MIT Press, 1481–1488.

Wainwright, M.J. and Jordan, M.I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learning* **1**, 1–305.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput.* **36**, 45–54. MR2370888

Wang, B. and Titterington, D.M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Anal.* **1**, 625–649. MR2221291

Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in MRFs. In *Advanced mean field methods: theory and practice*, ed. M. Opper and D. Saad, Cambridge, MA: MIT Press, 229–239. MR1863214