MAD-Bayes: MAP-based Asymptotic Derivations from Bayes

Tamara Broderick TAB@STAT.BERKELEY.EDU
UC Berkeley, Statistics Department

Brian Kulis KULIS@CSE.OHIO-STATE.EDU
Ohio State University, CSE Department

Michael I. Jordan JORDAN@EECS.BERKELEY.EDU
UC Berkeley, Statistics Department and EECS Department

Abstract

The classical mixture of Gaussians model is related to K-means via *small-variance asymptotics*: as the covariances of the Gaussians tend to zero, the negative log-likelihood of the mixture of Gaussians model approaches the K-means objective, and the EM algorithm approaches the K-means algorithm. Kulis & Jordan (2012) used this observation to obtain a novel K-means-like algorithm from a Gibbs sampler for the Dirichlet process (DP) mixture. We instead consider applying small-variance asymptotics directly to the posterior in Bayesian nonparametric models. This framework is independent of any specific Bayesian inference algorithm, and it has the major advantage that it generalizes immediately to a range of models beyond the DP mixture. To illustrate, we apply our framework to the feature learning setting, where the beta process and Indian buffet process provide an appropriate Bayesian nonparametric prior. We obtain a novel objective function that goes beyond clustering to learn (and penalize new) groupings for which we relax the mutual exclusivity and exhaustivity assumptions of clustering. We demonstrate several other algorithms, all of which are scalable and simple to implement. Empirical results demonstrate the benefits of the new framework.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

1. Introduction

Clustering is a canonical learning problem and arguably the dominant application of unsupervised learning. Much of the popularity of clustering revolves around the K-means algorithm; its simplicity and scalability make it the preferred choice in many large-scale unsupervised learning problems—even though a wide variety of more flexible algorithms, including those from Bayesian nonparametrics, have been developed since the advent of K-means (Steinley, 2006; Jain, 2010). Indeed, Berkhin (2006) writes that K-means is "by far the most popular clustering tool used nowadays in scientific and industrial applications."

K-means does have several known drawbacks. For one, the K-means algorithm clusters data into mutually exclusive and exhaustive clusters, which may not always be the optimal or desired form of latent structure for a data set. For example, pictures on a photo-sharing website might each be described by multiple tags, or social network users might be described by multiple interests. In these examples, a *feature allocation* in which each data point can belong to any nonnegative integer number of groups—now called *features*—is a more appropriate description of the data (Griffiths & Ghahramani, 2006; Broderick et al., 2013a). Second, the K-means algorithm requires advance knowledge of the number of clusters, which may be unknown or grow with the number of data points in some applications. A vast literature exists just on how to choose a number of clusters using heuristics or extensions of K-means (Steinley, 2006; Jain, 2010). A recent algorithm called DP-means (Kulis & Jordan, 2012) provides another perspective on the choice of cluster cardinality. Recalling the small-variance asymptotic argument that takes the EM algorithm for mixtures of Gaussians and yields the K-means algorithm, the authors apply this argument to a Gibbs sampler for a

Dirichlet process (DP) mixture (Antoniak, 1974; Escobar, 1994; Escobar & West, 1995) and obtain a K-means-like algorithm that does not fix the number of clusters upfront.

Notably, this derivation of DP-means is specific to the choice of the sampling algorithm and is also not immediately amenable to the feature learning setting. In this paper, we provide a more general perspective on these small-variance asymptotics. We show that one can obtain the objective function for DP-means (independent of any algorithm) by applying asymptotics directly to the MAP estimation problem of a Gaussian mixture model with a Chinese Restaurant Process (CRP) prior (Blackwell & MacQueen, 1973; Aldous, 1985) on the latent clustering. The key is to express the posterior in terms of the exchangeable partition probability function (EPPF) of the CRP (Pitman, 1995).

A critical advantage of this more general view of small-variance asymptotics is that it provides a framework for extending beyond the DP mixture. The Bayesian nonparametric toolbox contains many models that may yield—via small-variance asymptotics—a range of new algorithms that to the best of our knowledge have not been discovered in the K-means literature. We thus view our major contribution as providing new directions for researchers working on K-means and related discrete optimization problems.

To highlight this generality, we show how the framework may be used in the feature learning setting. We take as our point of departure the beta process (BP) (Hjort, 1990; Thibaux & Jordan, 2007), which is the feature learning counterpart of the DP, and the Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2006), which is the feature learning counterpart of the CRP. We show how to express the corresponding MAP inference problem via an analogue of the EPPF that we refer to as an "exchangeable feature probability function" (EFPF) (Broderick et al., 2013b). Taking an asymptotic limit we obtain a novel objective function for feature learning, as well as a simple and scalable algorithm for learning features in a data set. The resulting algorithm, which we call *BP-means*, is similar to the DP-means algorithm, but allows each data point to be assigned to more than one feature. We also use our framework to derive several additional algorithms, including algorithms based on the Dirichlet-multinomial prior as well as extensions to the marginal MAP problem in which the cluster/feature means are integrated out. We compare our algorithms to existing Gibbs sampling methods as well as existing hard clustering methods in order to highlight the benefits of our approach.

2. MAP Asymptotics for Clusters

We begin with the problem setting of Kulis & Jordan (2012) but diverge in our treatment of the small-variance asymptotics. We consider a Bayesian nonparametric framework for generating data via a prior on clusterings and a likelihood that depends on the (random) clustering. Prior and likelihood yield a posterior distribution. A point estimate of the clustering (i.e., a hard clustering) may be achieved by choosing a clustering that maximizes the posterior; the result is a *maximum a posteriori* (MAP) estimate.

Consider a data set $x_1, \ldots, x_N$, where x_n is a D-component vector. Let K^+ denote the (random) number of clusters. Let z_{nk} equal one if data index n belongs to cluster k and 0 otherwise, so there is exactly one value of k for each n such that $z_{nk} = 1$. We can order the cluster labels k so that the first K^+ clusters are non-empty (i.e., $z_{nk} = 1$ for some n for each such k). Together K^+ and $z_{1:N,1:K^+}$ describe a clustering.

The Chinese restaurant process (CRP) (Blackwell & MacQueen, 1973; Aldous, 1985) gives a prior on K^+ and $z_{1:N,1:K^+}$ as follows. Let $\theta > 0$ be a hyperparameter of the model. The first customer (data index 1) starts a new table in the restaurant; i.e., $z_{1,1} = 1$. Recursively, the nth customer (data index n) sits at an existing table k with probability in proportion to the number of people sitting there (i.e., in proportion to $S_{n-1,k} := \sum_{m=1}^{n-1} z_{mk}$) and at a new table with probability proportional to θ.

Suppose the final restaurant has K^+ tables with N total customers sitting according to $z_{1:N,1:K^+}$. Then the probability of this clustering is found from the above recursion:

$$\mathbb{P}(z_{1:N,1:K^+}) = \theta^{K^+-1} \frac{\Gamma(\theta+1)}{\Gamma(\theta+N)} \prod_{k=1}^{K^+} (S_{N,k}-1)!, \quad (1)$$

a formula that is known as an exchangeable partition probability function (EPPF) (Pitman, 1995).

A common choice for the likelihood is to assume that data in cluster k are Gaussian with cluster-specific mean μ_k and shared variance $\sigma^2 I_D$ (where I_D is the $D \times D$ identity matrix and $\sigma^2 > 0$). Then the likelihood of data $x = x_{1:N}$ given clustering $z = z_{1:N,1:K^+}$ and means $\mu = \mu_{1:K^+}$ is:

$$\mathbb{P}(x|z,\mu) = \prod_{k=1}^{K^+} \prod_{n:z_{n,k}=1} \mathcal{N}(x_n|\mu_k, \sigma^2 I_D).$$

Further suppose the μ_k are drawn iid Gaussian from a prior with mean 0 in every dimension and vari-

ance $\rho^2 I_D$ for hyperparameter $\rho^2 > 0$: $\mathbb{P}(\mu_{1:K^+}) = \prod_{k=1}^{K^+} \mathcal{N}(\mu_k|0, \rho^2 I_D)$.

The posterior distribution over the clustering given the observed data, $\mathbb{P}(z, \mu|x)$, is calculated from the prior and likelihood using Bayes theorem: $\mathbb{P}(z, \mu|x) \propto \mathbb{P}(x|z, \mu)\mathbb{P}(\mu)\mathbb{P}(z)$. We find the MAP point estimate for the clustering and cluster means by maximizing the posterior: $\text{argmax}_{K^+,z,\mu} \mathbb{P}(z, \mu|x)$. Note that the point estimate will be the same if we instead minimize the negative log joint likelihood: $\text{argmin}_{K^+,z,\mu} - \log \mathbb{P}(z, \mu, x)$.

In general, calculating the posterior or MAP estimate is difficult and usually requires approximation, e.g. via Markov chain Monte Carlo or a variational method. A different approximation can be obtained by taking the limit of the objective function above as the cluster variances decrease to zero: $\sigma^2 \to 0$. Since the prior allows an unbounded number of clusters, taking this limit will result in each data point being assigned to its own cluster in the MAP. To arrive at a limiting objective function that favors a non-trivial cluster assignment, we modulate the number of clusters via the hyperparameter θ, which varies linearly with the expected number of clusters in the prior. In particular, we choose some constant $\lambda^2 > 0$ and let $\theta = \exp(-\lambda^2/(2\sigma^2))$, so that, e.g., $\theta \to 0$ as $\sigma^2 \to 0$.

Substituting θ as a function of σ^2 and letting $\sigma^2 \to 0$, we find that $-2\sigma^2 \log \mathbb{P}(z, \mu, x)$ satisfies

$$\sim \sum_{k=1}^{K^+} \sum_{n:z_{nk}=1} \|x_n - \mu_k\|^2 + (K^+ - 1)\lambda^2, \qquad (2)$$

where $f(\sigma^2) \sim g(\sigma^2)$ here denotes $f(\sigma^2)/g(\sigma^2) \to 1$ as $\sigma^2 \to 0$. The double sum originates from the exponential function in the Gaussian data likelihood, and the penalty term—reminiscent of an AIC penalty (Akaike, 1974)—originates from the CRP prior (Sup. Mat. A).

From Eq. (2), we see that finding the MAP estimate of the CRP Gaussian mixture model is asymptotically equivalent to the following optimization problem:

$$\underset{K^+,z,\mu}{\text{argmin}} \sum_{k=1}^{K^+} \sum_{n:z_{nk}=1} \|x_n - \mu_k\|^2 + (K^+ - 1)\lambda^2. \qquad (3)$$

Kulis & Jordan (2012) derived a similar objective function, which they called the *DP-means objective function* (a name we retain for Eq. (3)), by first deriving a K-means-style algorithm from a DP Gibbs sampler. Here, by contrast, we have found this objective function directly from the MAP problem, with no reference to any particular inference algorithm and thereby demonstrating a more fundamental link between the MAP problem and Eq. (3). In the following, we show that this focus on limits of a MAP estimate can yield useful optimization problems in diverse domains.

Notably, the objective in Eq. (3) takes the form of the K-means objective function (the double sum) plus a penalty of λ^2 for each cluster after the first; this offset penalty is natural since any partition of a non-empty set must have at least one cluster.[1] Once we have Eq. (3), we may consider efficient solution methods; one candidate is the DP-means algorithm of Kulis & Jordan (2012).

3. MAP Asymptotics for Features

Once more consider a data set $x_{1:N}$, where x_n is a D-component vector. Now let K^+ denote the (random) number of features. Let z_{nk} equal one if data index n is in feature k and zero otherwise. In the feature case, while there must be a finite number of k values such that $z_{nk} = 1$ for any n, it is not required that there be exactly a single such k or even any such k. We order the feature labels k so that the first K^+ features are non-empty; i.e., we have $z_{nk} = 1$ for some n for each such k. Together K^+ and $z_{1:N,1:K^+}$ describe a feature allocation.

The Indian buffet process (IBP) (Griffiths & Ghahramani, 2006) is a prior on $z_{1:N,1:K^+}$ that places strictly positive probability on any finite, nonnegative value of K^+. Like the CRP, it is based on an analogy between the customers in a restaurant and the data indices. In the IBP, the dishes in the buffet correspond to features. Let $\gamma > 0$ be a hyperparameter of the model. The first customer (data index 1) samples $K_1^+ \sim \text{Pois}(\gamma)$ dishes from the buffet. Recursively, when the nth customer (data index n) arrives at the buffet, $\sum_{m=1}^{n-1} K_m^+$ dishes have been sampled by the previous customers. Suppose dish k of these dishes has been sampled $S_{n-1,k}$ times by the first $n-1$ customers. The nth customer samples dish k with probability $S_{n-1,k}/n$. The nth customer also samples $K_n^+ \sim \text{Pois}(\gamma/n)$ new dishes.

Suppose the buffet has been visited by N customers who sampled a total of K^+ dishes. Let $z = z_{1:N,1:K^+}$ represent the resulting feature allocation. Let H be the number of unique values of the $z_{1:N,k}$ vector across k; let $\tilde{K}_h$ be the number of k with the hth unique value of this vector. We calculate an "exchangeable feature probability function" (EFPF) (Broderick et al., 2013b) by multiplying together the probabilities from the N steps in the description and find that $\mathbb{P}(z)$ equals

[1] The objective of Kulis & Jordan (2012) penalizes all K^+ clusters; the optimal arguments are the same in each case.

(Griffiths & Ghahramani, 2006)

$$\frac{\gamma^{K^+}\exp\left\{-\sum_{n=1}^{N}\frac{\gamma}{n}\right\}}{\prod_{h=1}^{H}\tilde{K}_h!}\prod_{k=1}^{K^+}S_{N,k}^{-1}\binom{N}{S_{N,k}}^{-1}. \quad (4)$$

It remains to specify a probability for the observed data x given the latent feature allocation z. The linear Gaussian model of Griffiths & Ghahramani (2006) is a natural extension of the Gaussian mixture model to the feature case. As previously, we specify a prior on feature means $\mu_k \overset{iid}{\sim} \mathcal{N}(0, \rho^2 I_D)$ for some hyperparameter $\rho^2 > 0$. Now data point n is drawn independently with mean equal to the sum of its feature means, $\sum_{k=1}^{K^+} z_{nk}\mu_k$, and variance $\sigma^2 I_D$ for some hyperparameter $\sigma^2 > 0$. In the case where each data point belongs to exactly one feature, this model is just a Gaussian mixture. We often write the means as a $K \times D$ matrix A with kth row μ_k. Writing Z for the $N \times K$ matrix with (n,k) element z_{nk} and X for the $N \times D$ matrix with nth row x_n, we have $\mathbb{P}(X|Z,A)$ equal to

$$\frac{1}{(2\pi\sigma^2)^{ND/2}}\exp\left\{-\frac{\mathbf{tr}((X-ZA)'(X-ZA))}{2\sigma^2}\right\}. \quad (5)$$

As in the clustering case, we wish to find the joint MAP estimate of the structural component Z and group-specific parameters A. It is equivalent to find the values of Z and A that minimize $-\log\mathbb{P}(X,Z,A)$. Finally, we wish to take the limit of this objective as $\sigma^2 \to 0$. Lest every data point be assigned to its own separate feature, we modulate the number of features in the small-σ^2 limit by choosing some constant $\lambda^2 > 0$ and setting $\gamma = \exp(-\lambda^2/(2\sigma^2))$.

Letting $\sigma^2 \to 0$, we find that asymptotically (Sup. Mat. B)

$$-2\sigma^2\log\mathbb{P}(X,Z,A) \sim \mathbf{tr}[(X-ZA)'(X-ZA)] + K^+\lambda^2,$$

The trace originates from the matrix Gaussian, and the penalty term originates from the IBP prior.

It follows that finding the MAP estimate for the feature learning problem is asymptotically equivalent to solving:

$$\underset{K^+,Z,A}{\operatorname{argmin}}\ \mathbf{tr}[(X-ZA)'(X-ZA)] + K^+\lambda^2. \quad (6)$$

We follow Kulis & Jordan (2012) in referring to the underlying random measure when naming objective functions derived from Bayesian nonparametric priors. Recalling that the beta process (BP) (Hjort, 1990; Thibaux & Jordan, 2007) is the random measure underlying the IBP, we call the objective in Eq. (6) the *BP-means objective*. The trace term in Eq. (6) forms a K-means-style objective on a feature matrix Z and feature means A when the number of features (i.e., the number of columns of Z or rows of A) is fixed. The second term enforces a penalty of λ^2 for each feature. In contrast to the DP-means objective, even the first feature is penalized since $K^+ = 0$ is allowed here.

Algorithm 1 BP-means.

Iterate until no changes are made:

1. For $n = 1, \ldots, N$
   - For $k = 1, \ldots, K^+$, choose the optimal value (0 or 1) of z_{nk}.
   - Let Z' equal Z but with one new feature (labeled $K^+ + 1$) containing only data index n. Set $A' = A$ but with one new row: $A'_{K^+ + 1, \cdot} \leftarrow X_{n,\cdot} - Z_{n,\cdot}A$.
   - If the triplet $(K^+ + 1, Z', A')$ lowers the objective from the triplet (K^+, Z, A), replace the latter triplet with the former.
2. Set $A \leftarrow (Z'Z)^{-1}Z'X$.

We formulate a *BP-means algorithm* to solve the optimization problem in Eq. (6) and discuss its convergence properties. In Alg. 1, note that $Z'Z$ is invertible so long as no two features have the same collection of indices. If that is not the case, we simply combine the two features into a single feature before performing the inversion.

Proposition 1. *The BP-means algorithm converges after a finite number of iterations to a local minimum of the BP-means objective in Eq. (6).*

See Sup. Mat. G for the proof. Though the proposition guarantees convergence, it does not guarantee convergence to the global optimum—an analogous result to those available for the K-means and DP-means algorithms (Kulis & Jordan, 2012). Many authors have noted the problem of local optima in the clustering literature (Steinley, 2006; Jain, 2010). One expects that the issue of local optima is only exacerbated in the feature domain, where the combinatorial landscape is much more complex. In clustering, this issue is often addressed by multiple random restarts and careful choice of cluster initialization; in Section 5 below, we also make use of random algorithm restarts and propose a feature initialization akin to one with provable guarantees for K-means clustering (Arthur & Vassilvitskii, 2007).

4. Extensions

We demonstrate our methodology using different priors on Z below and using different likelihoods in Sup. Mat. F.

Collapsed objectives. It is believed that *collapsing* out the cluster or feature means from a Bayesian model by calculating instead the marginal structural posterior can improve MCMC sampler mixing in many scenarios (Liu, 1994). In the clustering case, collapsing translates to forming the posterior $\mathbb{P}(z|x) = \int_\mu \mathbb{P}(z,\mu|x)$. Note that even in the cluster case, we may use the matrix representations Z, X, and A so long as we make the additional assumption that $\sum_{k=1}^{K^+} z_{nk} = 1$ for each n. Finding the MAP estimate $\operatorname{argmax}_Z \mathbb{P}(Z|X)$ may, as usual, be accomplished by minimizing the negative log joint distribution with respect to Z. $\mathbb{P}(Z)$ is given by the CRP (Eq. (1)). $\mathbb{P}(X|Z)$ takes the form:

$$\frac{\exp\left\{-\frac{\mathbf{tr}\left(X'(I_N - Z(Z'Z+\frac{\sigma^2}{\rho^2}I_D)^{-1}Z')X\right)}{2\sigma^2}\right\}}{(2\pi\sigma^2)^{ND/2}(\rho^2/\sigma^2)^{K^+D/2}|Z'Z+\frac{\sigma^2}{\rho^2}I_D|^{D/2}}. \quad (7)$$

Eq. (7) was derived by Griffiths & Ghahramani (2006) for linear-Gaussian features but applies to Gaussian clusters when Z encodes a clustering. Using the same asymptotics in σ^2 and θ as before, we find the limiting optimization problem (Sup. Mat. C):

$$\operatorname*{argmin}_{K^+,Z} \mathbf{tr}(X'(I_N - Z(Z'Z)^{-1}Z')X) + (K^+ - 1)\lambda^2. \quad (8)$$

The first term in this objective was proposed, via independent considerations, by Gordon & Henderson (1977).

Simple algebraic manipulations allow us to rewrite the objective in a more intuitive format (Sup. Mat. C.1):

$$\operatorname*{argmin}_{K^+,Z} \sum_{k=1}^{K^+} \sum_{n:z_{nk}=1} \|x_{n,\cdot} - \bar{x}^{(k)}\|_2^2 + (K^+ - 1)\lambda^2, \quad (9)$$

where $\bar{x}^{(k)} := S_{N,k}^{-1} \sum_{m:z_{mk}=1} x_{m,\cdot}$ is the kth empirical cluster mean, i.e., the mean of all data points assigned to cluster k. This *collapsed DP-means objective* is just the original DP-means objective in Eq. (3) with the cluster means replaced by empirical cluster means. A corresponding optimization algorithm appears in Alg. 2. A similar proof to that of Kulis & Jordan (2012) shows that this algorithm converges in a finite number of iterations to a local minimum of the objective.

We have already noted that the likelihood associated with the Gaussian mixture model conditioned on a clustering is just a special case of the linear Gaussian model conditioned on a feature matrix. Therefore, it is not surprising that Eq. (7) also describes $\mathbb{P}(X|Z)$ when Z is a feature matrix. Now, $\mathbb{P}(Z)$ is given by the IBP (Eq. (4)). Using the same asymptotics in σ^2 and γ as in the joint MAP case, the MAP problem for feature allocation Z asymptotically becomes (Sup. Mat. D):

$$\operatorname*{argmin}_{K^+,Z} \mathbf{tr}(X'(I_N - Z(Z'Z)^{-1}Z')X) + K^+\lambda^2. \quad (10)$$

The key difference with Eq. (8) is that here Z may have any finite number of ones in each row. We call the objective in Eq. (10) the *collapsed BP-means objective*.

Just as the collapsed DP-means objective has an empirical cluster means interpretation, so does the collapsed BP-means objective have an interpretation in which the feature means matrix A in Eq. (6) is replaced by its empirical estimate $(Z'Z)^{-1}ZX$ (cf. Sup. Mat. G). In particular, we can rewrite the objective in Eq. (10) as $\mathbf{tr}[(X - Z(Z'Z)^{-1}Z'X)'(X - Z(Z'Z)^{-1}Z'X)] + K^+\lambda^2$. A corresponding optimization algorithm appears in Alg. 3. A similar proof to that of Proposition 1 shows that this algorithm converges in a finite number of iterations to a local minimum of the objective.

Parametric objectives. The generative models studied so far are *nonparametric* in the usual Bayesian sense; there is no a priori bound on the number of cluster or feature parameters. The objectives above are similarly nonparametric. Parametric models, with a fixed bound on the number of clusters or features, are often useful as well. See Sup. Mat. E for derivations

Algorithm 2 Collapsed DP-means.

Iterate until no changes are made:
1. For $n = 1, \ldots, N$
   - Assign x_n to the closest cluster if the contribution to the objective in Eq. (9) from the squared distance is at most λ^2.
   - Otherwise, form a new cluster with just x_n.

Algorithm 3 Collapsed BP-means.

Repeat the following step until no changes are made:
1. For $n = 1, \ldots, N$
   - Choose $z_{n,1:K^+}$ to minimize the objective in Eq. (10). Delete any redundant features.
   - Add a new feature (indexed $K^+ + 1$) with only data index n if doing so decreases the objective and if the feature would not be redundant.

Algorithm 4 K-features.

Repeat until no changes are made:
1. For $n = 1, \ldots, N$
   - For $k = 1, \ldots, K$, set $z_{n,k}$ to minimize $\|x_{n,1:K} - z_{n,1:K}A\|^2$.
2. Set $A = (Z'Z)^{-1}Z'X$.

of objectives for clustering and feature learning in the parametric case. Since below we apply the parametric version for the feature learning setting, which we call *K-features* (analogous to K-means but for feature learning), we include its description in Alg. 4.

5. Experiments

We examine collections of unlabeled data to discover latent shared features. We have already seen the BP-means and collapsed BP-means algorithms when the number of features is unknown. A third algorithm that we evaluate here involves running the K-features algorithm for different values of K and choosing the joint values of K, Z, A that minimize the BP-means objective in Eq. (6); we call this the *stepwise K-features algorithm*. If we assume the plot of the minimized K-features objective (Eq. (14)) as a function of K has increasing increments (i.e., decreasing negative increments), then we need only run the K-features algorithm for increasing K until the objective increases.

It is well known that the K-means algorithm is sensitive to the choice of cluster initialization (Peña et al., 1999). Potential methods of addressing this issue include multiple random initializations and choosing initial, random cluster centers according to the K-means++ algorithm (Arthur & Vassilvitskii, 2007). In the style of K-means++, we introduce a similar feature means initialization.

We first consider fixed K. In K-means++, the initial cluster center is chosen uniformly at random from the data set. However, we note that empirically, the various feature algorithms discussed tend to prefer the creation of a *base feature*, shared amongst all the data. So start by assigning every data index to the first feature, and let the first feature mean be the mean of all the data points. Recursively, for feature k with $k > 1$, calculate the distance from each data point $x_{n,\cdot}$ to its feature representation $z_{n,\cdot}A$ for the construction thus far. Choose a data index n with probability proportional to this distance squared. Assign $A_{k,\cdot}$ to be the nth distance. Assign $z_{m,k}$ for all $m = 1, \ldots, N$ to optimize the K-features objective. In the case where K is not known in advance, we repeat the recursive step as long as doing so decreases the objective.

Another important consideration in running these algorithms without a fixed number of clusters or features is choosing the relative penalty effect λ^2. One option is to solve for λ^2 from a proposed K value via a heuristic (Kulis & Jordan, 2012) or validation on a data subset. Rather than assume K and return to it in this roundabout way, in the following we aim merely to demonstrate that there exist reasonable values of λ^2 that return meaningful results. More carefully examining the translation from a discrete (K) to continuous (λ^2) parameter space may be a promising direction for future work.

Tabletop data. Using a LogiTech digital webcam, Griffiths & Ghahramani (2006) took 100 pictures of four objects (a prehistoric handaxe, a Klein bottle, a cellular phone, and a $20 bill) placed on a tabletop. The images are in JPEG format with 240 pixel height, 320 pixel width, and 3 color channels. Each object may or may not appear in a given picture; the experimenters endeavored to place each object (by hand) in a respective fixed location across pictures.

This setup lends itself naturally to the feature allocation domain. We expect to find a base feature depicting the tabletop and four more features, respectively corresponding to each of the four distinct objects. Conversely, clustering on this data set would yield either a cluster for each distinct feature combination—a much less parsimonious and less informative representation than the feature allocation—or some averages over feature combinations. The latter case again fails to capture the combinatorial nature of the data.

We emphasize a further point about identifiability within this combinatorial structure. One "true" feature allocation for this data is the one described above. But an equally valid allocation, from a combinatorial perspective, is one in which the base feature contains all four objects and the tabletop. There are four further features, each of which deletes an object and replaces it with tabletop so that every possible combination of objects on the tabletop can be constructed from the features. Indeed, any combination of objects on the tabletop could equally well serve as a base feature; the four remaining features serve to add or delete objects as necessary.

We run PCA on the data and keep the first $D = 100$ principal components to form the data vector for each image. This pre-processing is the same as that performed by Griffiths & Ghahramani (2006), except the authors in that case first average the three color channels of the images.

We consider the Gibbs sampling algorithm of Griffiths

Alg	Per run	Total	#
Gibbs	$8.5 \cdot 10^3$	—	10
Collap	11	$1.1 \cdot 10^4$	5
BP-m	0.36	$3.6 \cdot 10^2$	6
FeatK	0.10	$1.55 \cdot 10^2$	5

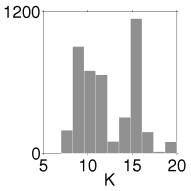

Figure 1. Left: A comparison of results for the IBP Gibbs sampler (Griffiths & Ghahramani, 2006), the collapsed BP-means algorithm, the basic BP-means algorithm, and the stepwise K-features algorithm. The first column shows the time for each run of the algorithm in seconds; the second column shows the total running time of the algorithm (i.e., over multiple repeated runs for the final three); and the third column shows the final number of features learned (the IBP # is stable for > 900 final iterations). *Right*: A histogram of collections of the final K values found by the IBP for a variety of initializations and parameter starting values.

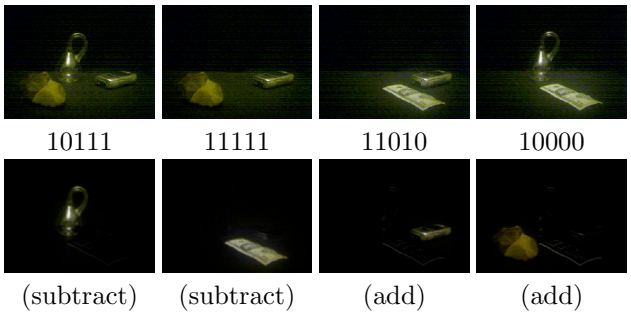

Figure 2. Upper row: Four example images in the tabletop data set. *Second row*: Feature assignments of each image. The first feature is the base feature, which depicts the Klein bottle and $20 bill on a tabletop and is almost identical to the fourth picture in the first row. The remaining four features are shown in order in the *third row*. The *fourth row* indicates whether the picture is added or subtracted when the feature is present.

& Ghahramani (2006) with initialization (mass parameter 1 and feature mean variance 0.5) and number of sampling steps (1000) determined by the authors; we explore alternative initializations below. We compare to the three feature means algorithms described above—all with $\lambda^2 = 1$. Each of the final three algorithms uses the appropriate variant of greedy initialization analogous to K-means++. We run 1000 random initializations of the collapsed and BP-means algorithms to mitigate local minima. We run 300 random initializations of K-features for each value of K and note that $K = 2, \ldots, 6$ are (dynamically) explored by the algorithm. All code was run in Matlab on the same computer. Timing and feature count results are on the left of Fig. 1.

While it is notoriously difficult to compare computation times for deterministic, hard-assignment algorithms such K-means to stochastic algorithms such as Gibbs sampling, particularly given the practical need for reinitialization to avoid local minima in the former, and difficult-to-assess convergence in the latter, it should be clear from the first column in the left-hand table of Fig. 1 that there is a major difference in computation time between Gibbs sampling and the new algorithms. Even when the BP-means algorithm is run 1000 times in a reinitialization procedure, the total time consumed is still an order of magnitude less than that for a single run of Gibbs sampling. Stepwise K-features is the fastest of the new algorithms.

We further note that if we were to take advantage of parallelism, additional drastic advantages could be obtained for the new algorithms. The Gibbs sampler requires each Gibbs iteration to be performed sequentially whereas the random initializations of the various feature means algorithms can be performed in parallel. A certain level of parallelism may even be exploited for the steps within each iteration of the collapsed and BP-means algorithms while the $z_{n,1:K}$ optimizations of K-features may all be performed in parallel across n

Another difficulty in comparing algorithms is that there is no clear single criterion with which to measure accuracy of the final model in unsupervised learning problems such as these. We do note, however, that theoretical considerations suggest that the IBP is not designed to find either a fixed number of features as N varies nor roughly equal sizes in those features it does find (Broderick et al., 2012). This observation may help explain the distribution of observed feature counts over a variety of IBP runs with the given data. To obtain feature counts from the IBP, we tried running in a variety of different scenarios—combining different initializations (one shared feature, 5 random features, 10 random features, initialization with the BP-means result) and different starting parameter values[2] (mass parameter values ranging logarithmically from 0.01 to 1 and mean-noise parameter values ranging logarithmically from 0.1 to 10). The final 100 K draws for each of these combinations are aggregated and summarized in a histogram on the right of Fig. 1. Feature counts lower than 7 were not obtained in our experiments, which suggests these values are, at least, difficult to obtain using the IBP with the given hyperpriors.

On the other hand, the feature counts for the new K-means-style algorithms suggest parsimony is more easily achieved in this case. The lower picture and text rows of Fig. 2 show the features (after the base fea-

[2]We found convergence failed for some parameter initializations outside this range.

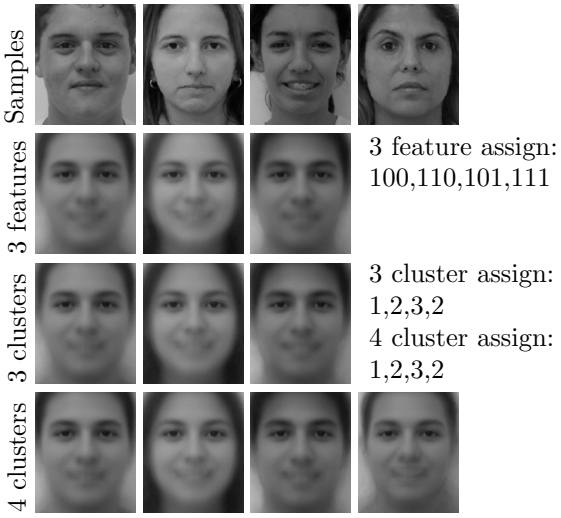

Figure 3. 1st row: Four sample faces. *2nd row*: The base feature (left) and other 2 features returned by stepwise K-features with $\lambda^2 = 5$. The final pictures are the cluster means from K-means with $K = 3$ (*3rd row*) and $K = 4$ (*4th row*). The righthand text shows how the sample pictures are assigned to features/clusters by each algorithm.

ture) found by stepwise K-features: as desired, there is one feature per tabletop object. The upper text row of Fig. 2 shows the features to which each of the example images in the top row are assigned by the optimal feature allocation. For comparison, the collapsed algorithm also finds an optimal feature encoding. The BP-means algorithm adds an extra, superfluous feature containing both the Klein bottle and $20 bill.

Faces data. Next, we analyze the FEI face database, consisting of 400 pre-aligned images of faces (Thomaz & Giraldi, 2010). 200 different individuals are pictured, each with one smiling and one neutral expression. Each picture has height 300 pixels, width 250 pixels, and one grayscale channel. Four example pictures appear in the first row of Fig. 3. This time, we compare the stepwise K-features algorithm to classic K-means. We keep the top 100 principal components to form the data vectors for both algorithms.

Given $\lambda^2 = 5$, stepwise K-features chooses one base feature (lefthand picture in the second row of Fig. 3) plus two additional features as optimal; the central and righthand pictures in the second row of Fig. 3 depict the sum of the base feature plus the corresponding feature. The second feature codes for longer hair and a shorter chin relative to the base feature. The third feature codes for darker skin and slightly different facial features. The feature combinations of each picture in the first row appear in the first text row on the right; all four possible combinations are represented.

K-means with 2 clusters and K-features with 2 features both encode exactly 2 distinct, disjoint groups. For larger numbers of groups though, the two representations diverge. For instance, consider a 3-cluster model of the face data, which has the same number of parameters as the 3-feature model. The resulting cluster means appear in the third row of Fig. 3. While the cluster means appear similar to the feature means, the assignment of faces to clusters is quite different. The second righthand text row in Fig. 3 shows to which cluster each of the four first-row faces is assigned. The feature allocation of the fourth picture in the top row tells us that the subject has long hair and certain facial features, roughly, whereas the clustering tells us that the subject's hair is more dominant than facial structure in determining grouping. Globally, the counts of faces for clusters (1,2,3) are (154,151,95) while the counts of faces for feature combinations (100,110,101,111) are (139,106,80,75).

We might also consider a clustering of size 4 since there are 4 groups specified by the 3-feature model. The resulting cluster means are in the bottom row of Fig. 3, and the cluster assignments of the sample pictures are in the bottom, righthand text row. None of the sample pictures falls in cluster 4. Again, the groupings provided by the feature allocation and the clustering are quite different. Notably, the clustering has divided up the pictures with shorter hair into 3 separate clusters. In this case, the counts of faces for clusters (1,2,3,4) are (121,150,74,55). The feature allocation here seems to provide a sparser and more interpretable representation relative to both cluster cardinalities.

6. Conclusions

We have developed a general methodology for obtaining hard-assignment objective functions from Bayesian MAP problems. The key idea is to include the structural variables explicitly in the posterior using combinatorial functions such as the EPPF and the EFPF. We apply this methodology to a number of generative models for unsupervised learning, with particular emphasis on latent feature models. We show that the resulting algorithms are capable of modeling latent structure out of reach of clustering algorithms but are also much faster than existing feature allocation learners from Bayesian nonparametrics.

Acknowledgments

We thank Tom Griffiths for generously sharing his code. Our research has been supported by NSF Graduate Research and Berkeley Fellowships, NSF award IIS-1217433, and ONR award N00014-11-1-0688.

References

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Aldous, D. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII1983*, pp. 1–198, 1985.

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

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Berkhin, P. A survey of clustering data mining techniques. *Grouping Multidimensional Data*, pp. 25–71, 2006.

Blackwell, D. and MacQueen, J. B. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

Broderick, T., Jordan, M. I., and Pitman, J. Beta processes, stick-breaking, and power laws. *Bayesian Analysis*, 7:439–476, 2012.

Broderick, T., Jordan, M. I., and Pitman, J. Clusters and features from combinatorial stochastic processes. *Statistical Science, to appear*, 2013a. Arxiv preprint arXiv:1206.5862.

Broderick, T., Pitman, J., and Jordan, M. I. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis, to appear*, 2013b. ArXiv preprint arXiv:1301.6647.

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

Escobar, M. D. and West, M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, pp. 577–588, 1995.

Gordon, A. D. and Henderson, J. T. An algorithm for Euclidean sum of squares classification. *Biometrics*, pp. 355–362, 1977.

Griffiths, T. and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. In Weiss, Y., Schölkopf, B., and Platt, J. (eds.), *Advances in Neural Information Processing Systems 18*, pp. 475–482. MIT Press, Cambridge, MA, 2006.

Hjort, N. L. Nonparametric Bayes estimators based on beta processes in models for life history data. *Annals of Statistics*, 18(3):1259–1294, 1990.

Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

Kulis, B. and Jordan, M. I. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *Proceedings of the 23rd International Conference on Machine Learning*, 2012.

Liu, J. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89:958–966, 1994.

Peña, J. M., Lozano, J. A., and Larrañaga, P. An empirical comparison of four initialization methods for the K-Means algorithm. *Pattern Recognition Letters*, 20(10):1027–1040, 1999.

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

Steinley, D. K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34, 2006.

Sung, K. and Poggio, T. Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51, 1998.

Thibaux, R. and Jordan, M. I. Hierarchical beta processes and the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 11, 2007.

Thomaz, C. E. and Giraldi, G. A. A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, 28(6):902–913, June 2010. We use files `http://fei.edu.br/~cet/frontalimages_spatiallynormalized_partX.zip` with `X=1,2`.