Approximate Maximum Likelihood Parameter Estimation For The Multiplicative Mixture Model And Overlapping Clustering

Priyank Patel ppatel@ece.utexas.edu
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78712-0240, USA

Editor: Inderjit Dhillon

Abstract

Most parametric clustering algorithms in use today employ generative models that do not have a natural mechanism to give rise to overlapping clusters. The multiplicative mixture model has been recently proposed as a generative model that can naturally give rise to overlapping clusters. However, performing maximum likelihood parameter estimation for this model using a standard technique like expectation maximization is intractable. As a result, Monte Carlo algorithms have been developed to do parameter estimation. In contrast to these stochastic algorithms, we propose a complementary deterministic algorithm to perform approximate maximum likelihood parameter estimation in a tractable manner. We then derive an overlapping clustering algorithm that uses employs the multiplicative mixture model as a generative model.

Keywords: Multiplicative Mixture Model, Approximate Maximum Likelihood Parameter Estimation, Overlapping Clustering, Variational Methods, Exponential Families.

1. Introduction

One of the most important tasks in unsupervised machine learning is that of data clustering. Parametric generative models for data clustering have been popular in the research literature due to the interpretable nature of the chosen generative model structure as well as the model parameters.

One of the most popular parametric generative models for clustering is the additive mixture model. It consists of K parameterized random variables, also known as mixture components, one of which is chosen at a time to generate a single data point. The random variable from which the data point is drawn is chosen according to a discrete probability distribution over the K mixture components. One of the reasons why the additive mixture model is so popular is that parameter estimation for it can be done in a computationally efficient manner by the expectation maximization algorithm. Redner and Walker (1984) provide a comprehensive review of additive mixture models. For a more modern unification of various additive mixture model based clustering algorithms done using Bregman divergences, see to Banerjee et al. (2005b).

However, most generative model based clustering algorithms do not have an in-built mechanism to detect overlapping clusters. This is not surprising since the structure of most generative models can not give rise to overlapping clusters. Yet, a lot of datasets have inherently overlapping clusters and quite a few applications such as mining gene expression data or text data require the ability to perform overlapping clustering. It was noted by Banerjee et al. (2005a) that when generative models such as the additive mixture model that can not naturally model overlapping clusters are nonetheless used to cluster data with overlapping clusters, then the quality of clustering results is very poor. Thus there is a need for generative models that can inherently model overlapping clusters. The idea of multiplying probability distributions together to get sharp distributions as well as overlaps in the output space was introduced by Hinton (2002) who called the resulting model a product of experts.

A multiplicative mixture model is further development of the product of experts model. It is a parametric generative model which consists of K underlying random variables which act as the mixture components that generate the observed data points. Associated with each observed data point is a hidden binary mixture vector of length K that indicates which of the K probability density functions corresponding to the mixture components were multiplied together and suitably normalized to get the probability distribution that generated that data point. Each data point is independently generated in this manner. Note that here any of the K mixture components can participate in generating a data point as opposed to additive mixture models where only one of the K mixture components can participate in generating a data point. Thus the structure of the multiplicative mixture model has a natural mechanism for generating overlapping clusters.

However this enhanced modeling capability comes at a cost. Infering the model parameters from the observed data is significantly more difficult in multiplicative mixture models than in additive mixture models. This is because the multiplicative model will have a normalization term that may not be expressible in closed form. This will make basic inference tasks like calculating marginal and posterior distributions analytically intractable. In addition, the hidden variables in multiplicative mixture models will have 2^K possible configurations each. Thus even if the normalization term is expressible in closed form, calculating the marginal and the posterior distributions will involve a summation over 2^K configurations which is computationally intractable for large K. Doing the same in additive mixture models will invole summation over just K configurations which is computationally tractable.

The intractability of the parameter estimation problem for the product of experts model itself was recognized at the very outset in (Hinton, 2002) and a Monte Carlo method was constructed for approximate inference of model parameters. A more recent Monte Carlo method for parameter estimation in the multiplicative model was developed by Heller and Ghahramani (2007). Williams et al. (2002) consider the specific case of Gaussian mixtures in a multiplicative model and draw connections to topics in machine learning like Probabilistic Principal Components Analysis and Probabilistic Minor Components Analysis.

For performing inference and learning of model parameters, a complementary alternative to stochastic approximation methods such as the Monte Carlo methods are the variational methods of constructing deterministic approximations (Jordan et al., 1999; Wainwright and Jordan, 2003) for quantities of interest via the introduction of additional or variational variables. Here, we take a similar deterministic approximation approach.

The rest of this report is organized as follows. In Section 2, we review some basic material regarding the exponential families of random variables that is useful subsequently. In Section 3, we describe the multiplicative mixture model and the intractability of maximum likelihood learning of its parameters. After that, in Section 4, we develop a determinitic method to do approximate maximum likelihood parameter estimation for the multiplicative mixture model and then use this method to devise an overlapping clustering algorithm in Section 5.

2. Exponential Families Of Random Variables

We begin with a brief review of some basic facts, stated without proof, regarding the exponential families of random variables and their connections to concepts from convex analysis. For a thorough treatment of exponential families refer to Amari and Nagaoka (2001); Barndorff-Nielsen (1978) and for that of convex analysis refer to Hiriart-Urruty and Lemaréchal (2001); Rockafellar (1970).

2.1 Basic Definitions

Let $\mathfrak{X}$ be a random variable that takes values x in some sample space $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\sigma(\mathcal{X})$ be a σ-algebra over $\mathcal{X}$ so that $(\mathcal{X}, \sigma(\mathcal{X}))$ is a measurable space. In addition, let P be the probability measure of the random variable $\mathfrak{X}$ and R be a given σ-finite reference measure where both measures are defined over the measurable space $(\mathcal{X}, \sigma(\mathcal{X}))$.

Further, let $t(x) = \{t_i(x)\}_{i=1}^s$ be a collection of measurable functions $t_i \colon \mathcal{X} \to \mathbb{R}$ that form the set of sufficient statistics for the random variable $\mathfrak{X}$. Thus, $t \colon \mathcal{X} \to \mathbb{R}^s$ is itself a vector valued measurable function. Also, let $\Theta \subseteq \mathbb{R}^s$ be defined as

$$\Theta = \left\{ \theta \in \mathbb{R}^s \,\middle|\, \int_{\mathcal{X}} e^{\langle \theta, t(x) \rangle} dR < \infty \right\},$$

where $\langle \theta, t(x) \rangle$ denotes the inner product in $\mathbb{R}^s$ between θ and $t(x)$. Based on this defintion of Θ we can define a function $\Upsilon \colon \Theta \to \mathbb{R}$ as

$$\Upsilon(\theta) = \log\left(\int_{\mathcal{X}} e^{\langle \theta, t(x) \rangle} dR \right).$$

Finally, given a $\theta \in \Theta$ we say that $\mathfrak{X}$ is a member of the exponential family of random variables that is parameterized by θ if its probability measure P is absolutely continuous with respect to the σ-finite reference measure R and the resulting Radon-Nikodym derivative of P with respect to R has the form

$$\frac{dP}{dR} = e^{\langle \theta, t(x) \rangle - \Upsilon(\theta)}.$$

The definitions of the set Θ and the function Υ ensure that P is a well defined and valid probability measure. The parameter θ is called the *natural parameter* and the set Θ is called the *natural parameter space*. The function Υ is called the *log-partition function* and it shall play an important role in the developments to follow. In some parts of the research literature, θ is also called the exponential parameter or the canonical parameter,

Θ is likewise called the exponential parameter space or the canonical parameter space, and Υ is also called the cumulant function.

If $\mathfrak{X}$ is a continuous random variable and the reference measure R is absolutely continuous with respect to the Lebesgue measure, then the probability measure P of $\mathfrak{X}$ is also absolutely continuous with respect to the Lebesgue measure and consequently has a Radon-Nikodym derivative that is of the form

$$\frac{dP}{dx} = e^{\langle \theta, t(x) \rangle - \Upsilon(\theta)} r(x),$$

where dx indicates the Radon-Nikodym derivative taken with respect to the Lebesgue measure and $r(x)$ is the Radon-Nikodym derivative of R with respect to the Lebesgue measure. This guarantees that there exists a probability density function $p(x;\theta) = \frac{dP}{dx}$ for $\mathfrak{X}$ and that $r \colon \mathcal{X} \to \mathbb{R}_+$ is a measurable function. Similarly, if $\mathfrak{X}$ is a discrete random variable and the reference measure R is absolutely continuous with respect to the counting measure, then in exactly the same manner as the continuous case, there exists a probability mass function $p(x;\theta) = \frac{dP}{dx}$ for $\mathfrak{X}$.

Henceforth, with a slight abuse of notation we will use dx to denote Radon-Nikodym derivatives taken with respect to both the Lebesgue and the counting measures, use the integral sign to denote both Lebesgue integrals and summations, and use the term probability density function to refer to probability density functions and probability mass functions for continuous and discrete random variables respectively. However, this should not cause any confusion since what is implied ought to be clear from the context.

Since all the continous and discrete exponential families of practical interest have probability measures that are absolutely continuous with respect to the Lebesgue and the counting measures respectively, we can effectively redefine the exponential families of random variables as those having a probability density function of the form

$$p(x;\theta) = e^{\langle \theta, t(x) \rangle - \Upsilon(\theta)} r(x), \tag{1}$$

where $r \colon \mathcal{X} \to \mathbb{R}_+$ is some positive function, $\theta \in \Theta$ which is defined as

$$\Theta = \left\{ \theta \in \mathbb{R}^s \middle| \int_{\mathcal{X}} e^{\langle \theta, t(x) \rangle} r(x) dx < \infty \right\},$$

and $\Upsilon \colon \Theta \to \mathbb{R}$ is defined as

$$\Upsilon(\theta) = \log\left(\int_{\mathcal{X}} e^{\langle \theta, t(x) \rangle} r(x) dx \right). \tag{2}$$

There are two important notions about an exponential family of random variables, those of a regular family and a minimal representation. These are defined next.

A Regular Family: An exponential family is said to be *regular* if the natural parameter space Θ is an open set. For regular exponential families, the log-partition function Υ is uniquely determined upto a constant additive factor.

A Minimal Representation: An exponential family is said to have a *minimal representation* if the components of the vector of sufficient statistics $t(x)$ are affinely independent almost everywhere, that is $\nexists$ non-zero $\alpha \in \mathbb{R}^s$ such that $\{\alpha, t(x)\} = \beta$, (a constant) $\forall x \in \mathcal{X}$

upto a set of measure zero. For exponential families with a minimal representation, $p(x;\theta)$ is uniquely determined for each natural parameter $\theta \in \Theta$.

Henceforth, we restrict our attention to regular exponential families with minimal representation due to the uniquely identifiable nature of the log-partition function and the natural parameter.

2.2 Properties of Θ and Υ

We next state some useful properties of the natural parameter space Θ and the log-partition function Υ of a regular exponential family of random variables with a minimal representation in terms of their connections to concepts from convex analysis.

Proposition 1 *The log-partition function $\Upsilon \colon \Theta \to \mathbb{R}$ is a convex function of Legendre type, that is, it satisfies the properties*

- *Θ is a non-empty and convex set.*
- *Υ is a proper, closed, strictly convex and differentiable function on Θ.*
- *$\|\nabla\Upsilon(\theta)\| \to \infty$ as θ approaches the boundary of Θ.*

As its alternate name implies, the log-partition function Υ is the cumulant generating function of the random vector $t(\mathfrak{X})$.

Proposition 2 *The log-partition function $\Upsilon \colon \Theta \to \mathbb{R}$ is a C^∞ function on Θ and is the cumulant generating function of the random vector $t(\mathfrak{X})$. In particular we have*

- *$\nabla\Upsilon(\theta) = \mathbb{E}[t(\mathfrak{X})]$.*
- *$\nabla^2\Upsilon(\theta) = \mathbb{E}[t(\mathfrak{X})t(\mathfrak{X})^T] - \mathbb{E}[t(\mathfrak{X})]\mathbb{E}[t(\mathfrak{X})]^T$.*

Since Υ is a C^∞ function on Θ, we are also guaranteed the existence of mixed cumulants of all orders for the random vector $t(\mathfrak{X})$.

2.3 Expectation And Conjugate Duality

Given a vector of sufficient statistics $t(x)$, the set of all $\mu \in \mathbb{R}^s$ such that μ is the expected value of $t(x)$ under some distribution that is absolutely continuous with respect to the reference measure R is called the *realizable mean parameter space M*. Note that M is a convex set.

From Proposition 2 we know that the gradient of the log-partition function $\nabla\Upsilon(\theta)$ equals the expected value of the the sufficient statistics of the exponential family distribution indexed by θ. The gradient is then a mapping from Θ onto the relative interior of M. This is somewhat surprising since the distribution in the definition of M was allowed to be an arbitrary one. As a result for a given μ in the relative interior of M we can restrict our attention to distributions of the exponential family.

In addition, the gradient mapping is one-to-one and hence invertible if and only if the exponential family has a minimal representation. The invertibility of the gradient mapping plays an important role in the conjugate function of the log-partition function and the relationship between the spaces Θ and M. We discuss this conjugate function next.

Family	$\mathcal{X}$	$r(x)$	dx	$t(x)$
Bernoulli	$\{0,1\}$	1	Counting	x
Binomial	$\{0,1,\ldots,N\}$	$\binom{N}{x}2^{-N}$	Counting	x
Poisson	$\{0,1,\ldots\}$	$\frac{1}{x!}$	Counting	x
Exponential	$\mathbb{R}_{++}$	1	Lebesgue	$-x$
Gaussian	$\mathbb{R}^d$	$\frac{1}{\sqrt{2\pi}^d}e^{-\frac{\|x\|^2}{2}}$	Lebesgue	x

Family	Θ	$\Upsilon(\theta)$	M	$\Upsilon^*(\mu)$
Bernoulli	$\mathbb{R}$	$\log(1+e^\theta)$	$[0,1]$	$\mu\log\mu+(1-\mu)\log(1-\mu)$
Binomial	$\mathbb{R}$	$N\log\left(\frac{1+e^\theta}{2}\right)$	$[0,N]$	$\mu\log\mu+(N-\mu)\log(N-\mu)+N\log\left(\frac{2}{N}\right)$
Poisson	$\mathbb{R}$	e^θ	$\mathbb{R}_+$	$\mu\log\mu-\mu$
Exponential	$\mathbb{R}_{++}$	$-\log\theta$	$\mathbb{R}_{--}$	$-\log(-\mu)-1$
Gaussian	$\mathbb{R}^d$	$\frac{1}{2}\|\theta\|^2$	$\mathbb{R}^d$	$\frac{1}{2}\|\mu\|^2$

Table 1: Examples Of Exponential Families.

Definition 3 *The conjugate function of $\Upsilon(\theta)$ is defined as*

$$\Upsilon^*(\mu)=\sup_{\theta\in\Theta}(\{\mu,\theta\}-\Upsilon(\theta)).$$

The supremum in this definition is achieved when $\mu=\nabla\Upsilon(\theta)$ and suppose μ lies in the relative interior of M. If the gradient mapping is invertible which it will be if the exponential family has a minimal representation, then the θ that achieves this supremum is the one that corresponds to the realizable mean μ and is given by $\theta=(\nabla\Upsilon)^{-1}(\mu)$.

Moreover, if the exponential family is regular in addition to having a minimal representation, then $\Upsilon(\theta)$ is a convex function of Legendre type and so is $\Upsilon^*(\mu)$. As a result $\Upsilon(\theta)$ can be written as the conjugate of it conjugate $\Upsilon^*(\mu)$ in the form

$$\Upsilon(\theta)=\sup_{\mu\in M}(\{\theta,\mu\}-\Upsilon^*(\mu)).$$

The supremum in the above representation is achieved when $\theta=\nabla\Upsilon^*(\mu)$. As $\Upsilon^*(\mu)$ is a convex function of Legendre type, it is strictly convex and differentiable and so $\nabla\Upsilon^*(\mu)$ is strictly monotonic and invertible. Hence the supremum is achieved by $\mu=(\nabla\Upsilon^*)^{-1}(\theta)$ or the realizable mean μ that corresponds to the natural parameter θ. But we also have the one-to-one gradient mapping. This gives us $(\nabla\Upsilon^*)^{-1}=\nabla\Upsilon$. To summarize,

Proposition 4 *If the exponential family is regular and has a minimal representation then*

- *$\Upsilon(\theta)$ and $\Upsilon^*(\mu)$ are convex functions of Legendre type and conjugates of one another.*
- *$\nabla\Upsilon(\theta)$ is a one-to-one mapping from Θ to M and $\nabla\Upsilon^*(\mu)$ is a one-to-one mapping from M to Θ and the two gradient mappings are inverses of each other.*

3. The Multiplicative Mixture Model

We now describe the multiplicative mixture model and the resulting intractability of doing maximum likelihood parameter estimation for it using a standard technique like expectation maximization.

3.1 Model Description

Let the K underlying random variables that form the mixture components of our model belong to the same known exponential family and $\hat{\theta} = \{\theta_k\}_{k=1}^K$ be the corresponding natural parameters. Thus $p(x;\theta_k)$ will be the probability density of the k^{th} mixture component. Also, let $x \in \mathcal{X}$ be an arbitrary observed data point, z be the corresponding binary mixture vector, and z_k be the k^{th} component of z. Each z_k is then associated with the k^{th} mixture component and its density $p(x;\theta_k)$. Under the multiplicative mixture model, the densities associated with those components of z that are set to 1 are chosen to be multiplied together and normalized to get the density according to which x is generated. Further, let $q(x)$ be some probability density that is also multiplied and normalized together with the chosen mixture component densities. We choose a suitable $q(x)$ to model background noise so that if none of the mixture components are chosen to generate x, then x is generated according to $q(x)$.

Implicit in this formulation is the assumption that the product of the chosen densitites and $q(x)$ is integrable so that it can be normalized to get a valid probability density. Unfortunately, the product of two or more densities is not guaranteed to be integrable. However, if we assume that $q(x)$ and $\{p(x;\theta_k)\}_{k=1}^K$ are bounded, then we are guaranteed that a product of $q(x)$ and any subset of the mixture component densitites is integrable . This is formalized in the following Theorem.

Theorem 5 *Let $q(x)$ and $\{p(x;\theta_k)\}_{k=1}^K$ be bounded probability densities. Then we have*

$$\int_{\mathcal{X}} q(x) \prod_{k \in \mathcal{K}} p(x;\theta_k)\, dx < \infty \quad \forall \mathcal{K} \subseteq \{1, 2, \ldots, K\}.$$

Proof See Appendix A. ■

Henceforth, we shall assume that the densities of the mixture components are bounded. This is an extremely mild assumption since the densities of all discrete exponential families will necessarily be bounded and those of all continuous exponential families of practical interest are bounded. In particular, note that all the densities of the commonly encountered exponential families listed in Table 1 are bounded.

Finally, let the K binary components of z be generated by K independent but not identically distributed Bernoulli random variables which, as described in Table 1, form a discrete exponential family. Let $\hat{\zeta} = \{\zeta_k\}_{k=1}^K$ then be the corresponding natural parameters with ζ_k being the natural parameter of the Bernoulli random variable that generates z_k. From the independence assumption we can express the joint density $p(z)$ of z as

$$\log p(z) = \sum_{k=1}^{K} \zeta_k z_k - \log(1 + e^{\zeta_k}). \tag{3}$$

Now, conditioned on a fixed configuration of z, we can express the conditional density $p(x|z)$ of x as

$$p(x|z) = \frac{q(x)\prod_{k=1}^{K} p(x;\theta_k)^{z_k}}{\int_{\mathcal{X}} q(x)\prod_{k=1}^{K} p(x;\theta_k)^{z_k}dx}.$$

From Theorem 5, this will be a well defined probability density. Taking a log on both sides of the previous expression and substituting exponential family densities of the form (1) into it, we get

$$\begin{aligned}\log p(x|z) &= \log q(x) + \sum_{k=1}^{K} z_k \log p(x;\theta_k) - \log\Bigg(\int_{\mathcal{X}} q(x)\prod_{k=1}^{K} p(x;\theta_k)^{z_k}dx\Bigg),\\ &= \log q(x) + \sum_{k=1}^{K} z_k\big[\langle\theta_k, t(x)\rangle + \log r(x)\big] - \log\Bigg(\int_{\mathcal{X}} q(x)\prod_{k=1}^{K}\big[e^{\langle\theta_k,t(x)\rangle}r(x)\big]^{z_k}dx\Bigg).\end{aligned} \tag{4}$$

3.2 Intractability Of Expectation Maximization

The multiplicative mixture model can be viewed more generally as a probabilistic model whose random variables are partitioned into two sets; one set consists of those variables whose realizations are observed and the other set consists of those variables whose realizations are hidden. In our case, the data point x is observed and the corresponding binary vector z that indicates which mixture components participated in generating x is hidden. For such models, expectation maximization (McLachlan and Krishnan, 1996) is a standard technique that is widely employed to do maximum likelihood parameter estimation. It consists of iteratively carrying out two successive steps known as the E-step and the M-step until some empirical criteria for convergence is met at which point the iterations are terminated and the algorithm stops. The incomplete log-likelihood function $\log p(x)$ is guaranteed to increase after each iteration and so we asymptotically converge to a stationary point of $\log p(x)$. For each such iteration of expectation maximization to be tractable, we need the E-step as well as the M-step to be tractable. Unfortunately, both steps are intractable for the multiplicative mixture model. To see why this is so, let us consider each step in turn.

In the E-step, we need to calculate the posterior density $p(z|x)$ of z conditioned on an observed x and to do this we will need to calculate the marginal density $p(x)$ of x. The intractability of doing so in evident from the integral term in (4). We encounter two problems here. First, the integral may not be expressible in closed form for our choice of $q(x)$ and the exponential family from which the mixture components are drawn as a result of which the E-step will become analytically intractable. Second, even if that integral was expressible in closed form, the K components of z will be coupled together in that expression in such a way that we will need to explicitly sum over all 2^K possible configurations of z in order to calculate $p(x)$ and hence $p(z|x)$. Thus, even if the E-step is analytically tractable, it will become computationally intractable.

In the M-step, we need to calculate the expectation of the complete log-likelihood function $\log p(x, z)$ with respect to the posterior density $p(z|x)$ that we get in the E-step and then maximize the resulting expected value over the model parameters $\hat{\theta}$ and $\hat{\zeta}$. Here, we again encounter the same problems as before. First, $\log p(x, z)$ may not be expressible in close form due to the integral term in (4) which will make the M-step analytically intractable. Second, even if it was expressible in closed form, the resulting expression will have have the K components of z coupled together in such a way that calculating the expectation of $\log p(x, z)$ with respect to any probability density over z, even a fully factorized one, will require us to explicitly sum over all 2^K possible configurations of z. After taking this expectation, we are still left with the problem of maximizing an objective function that consists of a sum of 2^K different terms. Hence, even if the M-step is analytically tractable, it will become computationally intractable.

4. Approximate Maximum Likelihood Parameter Estimation

As we saw in the last section, using expectation maximization to do exact maximum likelihood parameter estimation is intractable for the multiplicative mixture model. As an alternative, in this section we develop a tractable deterministic technique to do approximate instead of exact maximum likelihood parameter estimation. We do so by introducing additional free variables in a systematic manner in order to construct a closed form lower bound on the incomplete log-likelihood function $\log p(x)$. This lower bound will be in a form such that we can iteratively increase it in a tractable manner. However, this tractability does come at a cost. We now have additional free variables whose values must be determined and since we are increasing a lower bound on $\log p(x)$ at each iteration, our parameter estimation technique will converge to a stationary point of this lower bound. Unfortunately, that is not guaranteed to be a stationary point of $\log p(x)$ itself.

4.1 Constucting A Lower Bound On The Incomplete Log-Likelihood

We can get a closed form lower bound on the problematic integral in (4) by using the *perspective function* of the log-partition function Υ that is associated with the exponential family from which our mixture components are drawn. In addition, the components of z will become uncoupled in this lower bound which will let us efficiently marginalize it out later on. This lower bound is given in the following theorem.

Theorem 6 *Let $q(x) \leq 1$ and $r(x) \leq 1$. If $\hat{a} = \{a_k\}_{k=1}^K$ is such that each $a_k \in [0, 1]$ and $\sum_{k=1}^K a_k = 1$, then we have*

$$-\log\left(\int_{\mathcal{X}} q(x) \prod_{k=1}^K \left[e^{\langle \theta_k, t(x) \rangle} r(x)\right]^{z_k} dx\right) \geq -\sum_{k=1}^K z_k a_k \Upsilon\left(\frac{\theta_k}{a_k}\right), \tag{5}$$

where, for any $a_k = 0$ we define

$$a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) = \log\left(\operatorname*{ess\,sup}_{\mathcal{X}} e^{\langle \theta_k, t(x) \rangle} r(x)\right).$$

Proof See Appendix B. ■

The two conditions required for the lower bound in (5) to hold are rather mild. Since the choice of $q(x)$ is entirely upto us, we can always choose one that satifies $q(x) \leq 1$. Enforcing this condition might actually be desirable since we ought to choose a background noise distribution that is evenly spread out over the sample space and not concentrated in a few regions. Also, this condition will necessarily hold for a discrete $q(x)$. Similarly, the condition on $r(x)$ is also rather mild and is satisfied by most exponential families of practical interest. In particular, note that it is satisfied by all the commonly encountered families listed in Table 1. Moreover, if a given exponential family does not satisfy $r(x) \leq 1$, then we can always scale $r(x)$ by an appropriate constant factor and absorb its reciprocal into the log-partition function Υ.

We can now use the bound from Theorem 6 to construct a lower bound on the complete log-likelihood function $\log p(x, z)$. This lower bound will be in a form such that marginalization over the z present in it can be done in an efficient manner to give us a closed form lower bound on the incomplete log-likelihood function $\log p(x)$. It is this second bound that we are ultimately interested in for our approximate maximum likelihood parameter estimation problem. Let us begin by substituting (5) into (4) to get a lower bound on the conditional density $p(x|z)$ of x that has the form

$$\log p(x|z) \geq \log q(x) + \sum_{k=1}^{K} z_k \left[\langle \theta_k, t(x) \rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x) \right].$$

Next, adding (3) into the previous expression, we get a lower bound on the complete log-likelihood function $\log p(x, z)$ that can be expressed as

$$\log p(x, z) \geq \log q(x) + \sum_{k=1}^{K} z_k \left[\zeta_k + \langle \theta_k, t(x) \rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x) \right] - \log(1 + e^{\zeta_k}). \quad (6)$$

After exponentiating both sides, we get a lower bound on the joint density $p(x, z)$ of x and z that has the form

$$p(x, z) \geq q(x) \prod_{k=1}^{K} \frac{e^{z_k \left[\zeta_k + \langle \theta_k, t(x) \rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x) \right]}}{1 + e^{\zeta_k}}.$$

In the previous expression, we can see that the components of z have become uncoupled in the lower bound on the joint density. This will let us efficiently marginalize out the z to get a closed form lower bound on the density $p(x)$ of x in the following way.

$$\begin{aligned}
p(x) &= \sum_z p(x, z), \\
&\overset{(i)}{\geq} \sum_z q(x) \prod_{k=1}^K \frac{e^{z_k\left[\zeta_k + \langle\theta_k, t(x)\rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x)\right]}}{1 + e^{\zeta_k}}, \\
&= \sum_{z_1=0}^1 \sum_{z_2=0}^1 \cdots \sum_{z_K=0}^1 q(x) \prod_{k=1}^K \frac{e^{z_k\left[\zeta_k + \langle\theta_k, t(x)\rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x)\right]}}{1 + e^{\zeta_k}}, \\
&\overset{(ii)}{=} q(x) \prod_{k=1}^K \sum_{z_k=0}^1 \frac{e^{z_k\left[\zeta_k + \langle\theta_k, t(x)\rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x)\right]}}{1 + e^{\zeta_k}}, \\
&= q(x) \prod_{k=1}^K \frac{1 + e^{\left[\zeta_k + \langle\theta_k, t(x)\rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x)\right]}}{1 + e^{\zeta_k}},
\end{aligned}$$

where, the inequality in step (i) follows from the previous lower bound on $p(x, z)$ and exchanging the product and summations in step (ii) follows since each z_k appears in one and only one unique term within the product.

Taking a log on both sides, we get a lower bound on the incomplete log-likelihood function $\log p(x)$ that has the form

$$\log p(x) \geq \log q(x) + \sum_{k=1}^K \log\left(1 + e^{\left[\zeta_k + \langle\theta_k, t(x)\rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x)\right]}\right) - \log(1 + e^{\zeta_k}). \tag{7}$$

We now have a closed form lower bound on $\log p(x)$. Unfortunately, it has some undesirable properties. It is not concave in its variables and we would like to have a concave objective function in a maximization problem. Moreover, it is not even concave in any subset of its variables when the remaining ones are held fixed. In particular, note that if any two sets of variables out of $\{\hat{\theta}, \hat{\zeta}, \hat{a}\}$ are held fixed, then this lower bound is still not concave in the third set of variables. Thus, in order to iteratively increase (7), we will need to resort to some gradient ascent technique from non-linear programming. However, we do not need to do this. The lack of concavity is not as serious an obstacle as it might seem at first glance. We can exploit the variational characterization of the negative binary entropy function in terms of its convex conjugate to get a lower bound on (7) that can not only be made exact but also have the desired concavity properties. This lower bound is given in the following Lemma.

Lemma 7 *Let $h^*\colon \mathbb{R} \to \mathbb{R}_{++}$ be the function $h^*(y) = \log(1 + e^y)$ and $h\colon [0, 1] \to [-\log 2, 0]$ be the negative binary entropy function $h(b) = b \log b + (1 - b) \log(1 - b)$. Then h^* and h are convex conjugates of one another and for $y \in \mathbb{R}$, $b \in [0, 1]$, we will have*

$$h^*(y) = \max_{b\in[0,1]} \{by - h(b)\} \quad \textit{where} \quad \arg\max_{b\in[0,1]} \{by - h(b)\} = \frac{1}{1 + e^{-y}}, \tag{8}$$

$$h(b) = \max_{y\in\mathbb{R}} \{by - h^*(y)\} \quad \textit{where} \quad \arg\max_{y\in\mathbb{R}} \{by - h^*(y)\} = \log\left(\frac{b}{1 - b}\right), \tag{9}$$

and

$$h^*(y) + h(b) \geq by. \tag{10}$$

Proof See Appendix C. ■

We can now use the bound from Lemma 7 to construct a lower bound on $\log p(x)$. Let $\hat{b} = \{b_k\}_{k=1}^K$ be such that each $b_k \in [0, 1]$. Substituting (10) into (7) and using a unique b_k from $\hat{b}$ for each one of the K terms in the summation, we get a lower bound on the incomplete log-likelihood function that has the form

$$\log p(x) \geq \log q(x) + \sum_{k=1}^{K} b_k \left[\zeta_k + \langle \theta_k, t(x) \rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x) \right] - h(b_k) - \log(1 + e^{\zeta_k}). \tag{11}$$

We now have a closed form lower bound on $\log p(x)$ that has the desired concavity properties. While it is still not concave in all its variables, the bound in (11) is concave in any one set of variables out of $\{\hat{\theta}, \hat{\zeta}, \hat{a}, \hat{b}\}$ when the other three sets of variables are held fixed. Also, the set of all feasible values for each one of the four sets of variables is a convex set. This suggests a straightforward block co-ordinate ascent scheme to iteratively increase this lower bound since maximizing the objective function along any one block of co-ordinates will reduce to a somewhat simple convex program for most exponential families. Moreover, as we shall see later on, the particular form of this lower bound will lead to a very natural and intuitive cluster assignment rule.

4.2 Extending The Lower Bound To An Entire Dataset

So far, we have only considered a single observed data point x that is generated by the multiplicative mixture model and a lower bound on its incomplete log-likelihood function $\log p(x)$. Since our model generates all the observed data points in an independent and identical manner, we can construct a lower bound on the log-likelihood of an entire dataset in a rather straightforward fashion. Let $\mathcal{D}$ be a dataset consisting of N observed data points. A direct use of the lower bound in (11) will require us to introduce a new pair of the free variables $\{\hat{a}, \hat{b}\}$ for each one of the N data points. Doing this might be computationally too demanding for very large data sets. Hence, we will introduce the free variables in more controlled and scalable manner.

Let us split the dataset $\mathcal{D}$ into M partitions and let $\mathcal{D}_m$ be the m^{th} partition with N_m data points in it. Instead of introducing N_m pairs of free variables, one for each data point, we will introduce just one pair $\{\hat{a}_m, \hat{b}_m\}$ of free variables for all the data points in $\mathcal{D}_m$. In this way, by sharing free variables among the data points, we can control their number in the lower bound on the log-likelihood. However, this control does come at a cost. Reducing the number of free variables will result in a looser lower bound on the log-likelihood which will then adversely affect the quality of our approximate maximum likelihood parameter estimates. This is undesirable and so we should select a M as high as possible within the constraints of the available computing resources.

Let $\{\hat{a}_m, \hat{b}_m\} = \{a_{mk}, b_{mk}\}_{k=1}^K$ be the free variables for the m^{th} data partition $\mathcal{D}_m$. Then, we can get a lower bound on the incomplete log-likelihood function of all the data points in $\mathcal{D}_m$ by using the lower bound on $\log p(x)$ as follows.

$$\begin{aligned}\log p(\mathcal{D}_m) &= \sum_{x\in\mathcal{D}_m} \log p(x),\\ &\overset{(i)}{\geq} q_m + N_m \sum_{k=1}^{K} b_{mk}\left[\zeta_k + \langle\theta_k, t_m\rangle - a_{mk}\Upsilon\left(\frac{\theta_k}{a_{mk}}\right) + r_m\right] - h(b_{mk}) - \log(1+e^{\zeta_k}),\end{aligned} \tag{12}$$

where, the inequality in step (i) follows from the lower bound in (11) and from a summation over all the data points in $\mathcal{D}_m$ while holding $\{\hat{\theta}, \hat{\zeta}, \hat{a}_m, \hat{b}_m\}$ fixed. The constants q_m, t_m, and r_m depend upon the data partition $\mathcal{D}_m$ and are defined as

$$\begin{aligned} q_m &= \sum_{x\in\mathcal{D}_m} \log q(x),\\ t_m &= \frac{1}{N_m}\sum_{x\in\mathcal{D}_m} t(x),\\ r_m &= \frac{1}{N_m}\sum_{x\in\mathcal{D}_m} \log r(x).\end{aligned}$$

Finally, we can get a lower bound on the incomplete log-likelihood function of all the data points in our dataset $\mathcal{D}$ by using the lower bound on $\log p(\mathcal{D}_m)$ as follows.

$$\begin{aligned}\log p(\mathcal{D}) &= \sum_{m=1}^{M} \log p(\mathcal{D}_m),\\ &\overset{(i)}{\geq} q + \sum_{m=1}^{M} N_m \sum_{k=1}^{K} b_{mk}\left[\zeta_k + \langle\theta_k, t_m\rangle - a_{mk}\Upsilon\left(\frac{\theta_k}{a_{mk}}\right) + r_m\right] - h(b_{mk}) - \log(1+e^{\zeta_k}),\end{aligned} \tag{13}$$

where, the inequality in step (i) follows from the lower bound in (12) and from a summation over all the data partitions $\mathcal{D}_m$. The constant q depends upon all the data points in $\mathcal{D}$ and is defined as

$$q = \sum_{m=1}^{M} q_m.$$

Maximizing the lower bound in (13) will give us the approximate maximum likelihood estimates of the model parameters $\hat{\theta}$ and $\hat{\zeta}$ that we are ultimately interested in. In addition, the values of the free parameters at this maximum will also be of interest to us when we devise an overlapping clustering algorithm for the dataset $\mathcal{D}$.

4.3 Iteratively Increasing The Lower Bound

We now look at how to maximize the lower bound in (13) on the incomplete log-likelihood function $\log p(\mathcal{D})$ of the dataset $\mathcal{D}$. For the sake of notational convenience, let us define the function L to be this lower bound. In order to keep our notation uncluttered, we will suppress the dependence of L on the variables $\hat{\theta}$, $\hat{\zeta}$, $\{\hat{a}_m\}_{m=1}^M$, and $\{\hat{b}_m\}_{m=1}^M$ that define it. These variables will be left implicit whenever we refer to L. Thus we have

$$L = q + \sum_{m=1}^{M} N_m \sum_{k=1}^{K} b_{mk} \left[\zeta_k + \langle \theta_k, t_m \rangle - a_{mk} \Upsilon\left(\frac{\theta_k}{a_{mk}}\right) + r_m \right] - h(b_{mk}) - \log(1 + e^{\zeta_k}). \quad (14)$$

Since L is not a concave function, we can not expect to find its global maximum and so we will have to settle for a local maximum instead. Even though L may not be concave, it will be concave in any one set of variables out of $\hat{\theta}$, $\hat{\zeta}$, $\{\hat{a}_m\}_{m=1}^M$, and $\{\hat{b}_m\}_{m=1}^M$ when the other three sets of variables are held fixed. Also, the set of all feasible values for each one of the four sets of variables is a convex set. Quite unsurprisingly, L has analogous concavity properties to the lower bound in (11) on the log-likelihood of a single data point. As mentioned before, this suggests a natural block co-ordinate ascent scheme to iteratively increase L since maximizing it along any one block of co-ordinates will reduce to a relatively simple convex program for most choices of exponential families. Also, we can verify in a rather straighforward manner that L is a continuously differentiable function of all its variables.

Let us now turn our attention to maximizing L over any one out of its four sets of variables when the other three are held fixed. We begin with the natural parameters $\hat{\zeta}$ of the independent Bernoulli random variables that generate z. From the definition of L in (14), we will have

$$\begin{aligned}\max_{\hat{\zeta}} L &= \max_{\hat{\zeta}} \sum_{m=1}^{M} N_m \sum_{k=1}^{K} b_{mk}\zeta_k - \log(1 + e^{\zeta_k}),\\ &\overset{(i)}{=} \max_{\hat{\zeta}} \sum_{k=1}^{K} \left[\zeta_k \left(\sum_{m=1}^{M} \frac{N_m}{N} b_{mk} \right) - \log(1 + e^{\zeta_k}) \right],\end{aligned}$$

where, step (i) follows from exchanging the two summations and using the fact that $\sum_{m=1}^{M} N_m = N$ is a positive constant that we can divide by without changing where the maximum is attained. Since the objective function here decomposes into an uncoupled sum over the K components $\{\zeta_k\}_{k=1}^K$ of $\hat{\zeta}$, we can maximize it over $\hat{\zeta}$ by independently maximizing it over each individual ζ_k. Using result (9) from Lemma 7, we can express the solution to this maximization problem and thus the update rule for each one of the K components of $\hat{\zeta}$ as

$$\arg\max_{\zeta_k} L = \log\left(\frac{\sum_{m=1}^{M} N_m b_{mk}}{N - \sum_{m=1}^{M} N_m b_{mk}} \right). \quad (15)$$

Next, we look at maximizing L over the free variables $\{\hat{b}_m\}_{m=1}^M$. From its definition in (14), we can see that as a function of $\{\hat{b}_m\}_{m=1}^M$ when the other three sets of variables are held fixed, L will decompose into an uncoupled sum over the the M data partitions $\mathcal{D}_m$. Thus we can maximize L over $\{\hat{b}_m\}_{m=1}^M$ by independently maximizing it over each $\hat{b}_m$. This maximization can be expressed as

$$\max_{\hat{b}_m} L = \max_{\hat{b}_m} \sum_{k=1}^{K} b_{mk} \left[\zeta_k + \langle \theta_k, t_m \rangle - a_{mk} \Upsilon\left(\frac{\theta_k}{a_{mk}}\right) + r_m \right] - h(b_{mk}),$$

where, we have divided out the positive constant N_m as it will not change where the maximum is attained. Here as well, the objective function decomposes into an uncoupled sum over the K components $\{b_{mk}\}_{k=1}^{K}$ of $\hat{b}_m$. Hence we can maximize it over $\hat{b}_m$ by maximizing it individually over each b_{mk}. Using result (8) from Lemma 7, we can express the solution to this maximization problem and thus the update rule for each one of the K components of $\hat{b}_m$ in the form

$$\arg\max_{b_{mk}} L = \frac{1}{1 + e^{-\left[\zeta_k + \langle \theta_k, t_m \rangle - a_{mk} \Upsilon\left(\frac{\theta_k}{a_{mk}}\right) + r_m\right]}}. \tag{16}$$

Now, we look at maximizing L over the natural parameters $\hat{\theta}$ of the mixture components. From (14) we will get

$$\begin{aligned}\max_{\hat{\theta}} L &= \max_{\hat{\theta}} \sum_{m=1}^{M} N_m \sum_{k=1}^{K} b_{mk} \left[\langle \theta_k, t_m \rangle - a_{mk} \Upsilon\left(\frac{\theta_k}{a_{mk}}\right) \right], \\ &\overset{(i)}{=} \max_{\hat{\theta}} \sum_{k=1}^{K} \left[\left\langle \theta_k, \sum_{m=1}^{M} N_m b_{mk} t_m \right\rangle - \sum_{m=1}^{M} N_m b_{mk} a_{mk} \Upsilon\left(\frac{\theta_k}{a_{mk}}\right) \right],\end{aligned}$$

where step (i) follows from exchanging the two summations and using the linearity of an inner product. As before, we can see that the objective function will decompose into an uncoupled sum over the K components $\{\theta_k\}_{k=1}^{K}$ of $\hat{\theta}$ and so, we can maximize it over $\hat{\theta}$ by independently maximizing it over each individual θ_k. Unfortunately, unlike the two previous maximizations, we can not express the solution to this problem in closed form. This is because, the solution will heavily depend upon the nature of the log-partition function Υ and hence the choice of the exponential family from which the mixture components are drawn. Nonetheless this is a convex program and it will have a unique solution. We can still express the update rule for each one of the K components of $\hat{\theta}$ as

$$\arg\max_{\theta_k} L = \arg\max_{\theta_k} \left[\left\langle \theta_k, \sum_{m=1}^{M} N_m b_{mk} t_m \right\rangle - \sum_{m=1}^{M} N_m b_{mk} a_{mk} \Upsilon\left(\frac{\theta_k}{a_{mk}}\right) \right]. \tag{17}$$

Finally, we look at maximizing L over the free variables $\{\hat{a}_m\}_{m=1}^{M}$. From (14), we can see that L will decompose into an uncoupled sum over the the M data partitions $\mathcal{D}_m$ when $\{\hat{a}_m\}_{m=1}^{M}$ is allowed to vary and the other three sets of variables are held fixed. Thus we can maximize L over $\{\hat{a}_m\}_{m=1}^{M}$ by independently maximizing it over each $\hat{a}_m$. This maximization will be of the form

$$\max_{\hat{a}_m} L = \max_{\hat{a}_m} \sum_{k=1}^{K} -b_{mk} a_{mk} \Upsilon\left(\frac{\theta_k}{a_{mk}}\right),$$

where, we have divided out the positive constant N_m since it will not change the location of the maxima. Unlike the previous three maximizations, the objective function here will not decompose into an uncoupled sum over the K components $\{a_{mk}\}_{k=1}^{K}$ of $\hat{a}_m$. This is because of the $\sum_{k=1}^{K} a_{mk} = 1$ constraint from Theorem 6. Hence we will have to maximize it jointly over the a_{mk}. Moreover, we will not be able to express the solution to this maximization in closed form due to its heavy dependence on the nature of the log-partition function of our chosen exponential family. However, it is a convex program and will have a unique solution. We can still write the update rule for $\hat{a}_m$ as

$$\begin{aligned} \arg\max_{\hat{a}_m} L = \arg\max \sum_{k=1}^{K} -b_{mk} a_{mk} \Upsilon\left(\frac{\theta_k}{a_{mk}}\right), \\ \text{s.t.} \sum_{k=1}^{K} a_{mk} = 1,\ a_{mk} \geq 0. \end{aligned} \tag{18}$$

It is clear from the update rules for the variables that our objective function will increase after each iteration of the update cycle. Moreover, the concavity of L along each set of variables and the fact that it is continuously differentiable over all the variables will also guarantee that the sequence of iterates generated in this manner will converge to a local maximum of L. This is convergence guarantee formalized in the following lemma.

Lemma 8 *Every limit point of the sequence of iterates generated by the update cycle given in* (15)*,* (16)*,* (17)*, and* (18) *is a stationary point of L.*

Proof See Bertsekas (1999, Proposition 2.7.1). ■

5. An Overlapping Clustering Algorithm

We can use the approximate maximum likelihood parameter estimation scheme developed in the previous section to derive an overlapping clustering algorithm for which the multiplicative mixture model serves as a generative model. Our clustering algorithm will return a hard clustering of the dataset $\mathcal{D}$, that is, each data point will either be included in or excluded from each cluster. Each one of the K mixture components in our multiplicative model corresponds to one cluster and each data point can simultaneously belong to any subset of the these K clusters.

Our clustering algorithm proceeds in two distinct stages. First, we obtain approximate maximum likelihood estimates of the model parameters $\hat{\theta}$ and $\hat{\zeta}$ along the associated free variables $\{\hat{a}_m, \hat{b}_m\}_{m=1}^{M}$ using the block co-ordinate ascent scheme described in the previous section. Next, we infer the cluster assignments. The binary mixture vector z associated with each data point x encodes which mixture components participated in generating x, it will encode which clusters x belongs to. Since z is hidden from us, the most obvious and natural estimate of z given a particular x would be the one that had the maximum probability of occuring conditioned on the observed x. Thus the cluster assignment for the data point x will be given by the maximum aposteriori probability estimate of z that is given by

$$\arg\max_{z} p(z|x).$$

However, as we saw in Section 3, calculating $p(z|x)$ for a single z becomes intractable. Hence maximizing $p(z|x)$ over all 2^K possible configurations of z is clearly intractable as well. As an alternative to maximizing $p(z|x)$, we will instead maximize a lower bound on it. The form of this lower bound will be such that maximizing it over all configurations of z can be done in a tractable manner. This is in the same spirit as our approximate maximum likelihood estimation scheme, where instead of maximizing the log-likelihood of the dataset, we maximized a lower bound on it. We proceed as follows

$$\begin{aligned}\arg\max_z p(z|x) &= \arg\max_z \frac{p(x,z)}{p(x)},\\ &\overset{(i)}{=} \arg\max_z p(x,z),\\ &\overset{(ii)}{=} \arg\max_z \, \log p(x,z).\end{aligned}$$

where, step (i) follows since for a particular x, $p(x)$ is a positive constant independent of z and hence it will not change the z at which the maximum is achieved and step (ii) follows since the $\log(\cdot)$ function is strictly monotone increasing.

We have already derived a lower bound given in (6) on the complete log-likelihood function $\log p(x,z)$ of a particular data point x. We will maximize this lower bound over all possible configurations of z to get the cluster assignments for x. For the sake of notational convenience, let us define the function J to be this lower bound. Then, from (6), we will have

$$J = \log q(x) + \sum_{k=1}^{K} z_k \left[\zeta_k + \langle \theta_k, t(x)\rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x)\right] - \log(1+e^{\zeta_k}).$$

where, the free variables $\hat{a} = \{a_k\}_{k=1}^{K}$ and $\hat{b} = \{b_k\}_{k=1}^{K}$ are the ones associated with the data partition that x lies in and their values are taken to be those at the termination of the block co-ordinate ascent iterations.

Maximizing J over the z, we get

$$\arg\max_z J = \arg\max_z \sum_{k=1}^{K} z_k \left[\zeta_k + \langle \theta_k, t(x)\rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x)\right],$$

We can see that the objective function for this maximization problem has decomposed into an uncoupled sum over the K mixture components that correspond to the K clusters. Hence we can maximize J over all possible configurations of z by independently maximizing it over each individual component z_k of z. Thus maximizing z will reduce to individually maximizing $z_k J_k(x)$ over z_k, where the functions $J_k(x)$ are defined as

$$J_k(x) = \left[\zeta_k + \langle \theta_k, t(x)\rangle - a_k \Upsilon\left(\frac{\theta_k}{a_k}\right) + \log r(x)\right].$$

Since, each z_k can be either 0 or 1, $J_k(x)$ will be maximized by $z_k = 0$ if $J_k(x) \leq 0$ and conversely, $J_k(x)$ will be maximized by $z_k = 1$ if $J_k(x) > 0$. This assignment of z_k based on the sign of $J_k(x)$ will then be our cluster assignment rule. This has a very natural and intuitive interpretation. For every data point x, the model parameters $\hat{\theta}$ and $\hat{\zeta}$, along with the free variables $\hat{a}$ of the data partition that x lies in will together determine a decision region for each on the K clusters. The boundaries of these K decision regions will be given by $J_k(x) = 0$. If x lies inside the decision region for the k^{th} cluster, that is if $J_k(x) > 0$, then x is included in the k^{th} cluster. Otherwise, x is excluded from that cluster. This is done one after the other for the decision regions of all K clusters.

Acknowledgments

The author would like to thank Sreangsu Acharyya for valuable discussions and helpful comments.

Appendix A. Proof Of Theorem 5

In this appendix we prove Theorem 5 from Section 3. We begin with a basic result from real analysis that will be useful in the proof.

Proposition 9 (Tchebyshev's Inequality) *Let $f\colon \mathcal{X} \to \mathbb{R}_+$ be a non-negative and measurable function on $\mathcal{X} \subseteq \mathbb{R}^d$. If $\alpha > 0$, then we have*

$$|\{x \in \mathcal{X} \mid f(x) > \alpha\}| \leq \frac{1}{\alpha} \int_{\mathcal{X}} f(x) dx,$$

where $|\cdot|$ denotes the Lebesgue measure of a set.

Proof See Wheeden and Zygmund (1977, Corollary 5.12). ■

Next, we present a proof of Theorem 5. For the sake of readability, we repeat the statement of the theorem before presenting its proof.

Theorem *Let $q(x)$ and $\{p(x;\theta_k)\}_{k=1}^K$ be bounded probability densities. Then we have*

$$\int_{\mathcal{X}} q(x) \prod_{k \in \mathcal{K}} p(x;\theta_k)\, dx < \infty \quad \forall \mathcal{K} \subseteq \{1, 2, \ldots, K\}.$$

Proof Let us fix an arbitrary subset $\mathcal{K}$ of $\{1, 2, \ldots, K\}$ and define the following sets:

$$\begin{aligned} \mathcal{Y}_q &= \{x \in \mathcal{X} \mid q(x) > 1\}, \\ \mathcal{Y}_k &= \{x \in \mathcal{X} \mid p(x;\theta_k) > 1\}, \\ \mathcal{Y}_{\mathcal{K}} &= \mathcal{Y}_q \cup \bigcup_{k \in \mathcal{K}} \mathcal{Y}_k. \end{aligned}$$

Then, from Tchebyshev's inequality with $\alpha = 1$ and the subadditivity of the Lebesgue measure, we get

$$\begin{aligned}
|\mathcal{Y}_q| &\leq \int_{\mathcal{X}} q(x)dx = 1, \\
|\mathcal{Y}_k| &\leq \int_{\mathcal{X}} p(x;\theta_k)dx = 1, \\
|\mathcal{Y}_{\mathcal{K}}| &\leq |\mathcal{Y}_q| + \sum_{k\in\mathcal{K}} |\mathcal{Y}_k| \leq 1 + |\mathcal{K}|, \qquad (19)
\end{aligned}$$

where, with a slight abuse of notation, $|\mathcal{K}|$ denotes the number of elements in the set $\mathcal{K}$. Our use of $|\cdot|$ to denote the Lebesgue measure of a set as well as the number of elements in a finite set should not cause any confusion since which one is implied ought to be clear from the set in question.

From the defintion of $\mathcal{Y}_{\mathcal{K}}$ we can see that if $x \notin \mathcal{Y}_{\mathcal{K}}$, that is if $x \in \mathcal{X} \setminus \mathcal{Y}_{\mathcal{K}}$, then $q(x) \leq 1$ and $p(x;\theta_k) \leq 1$ for each $k \in \mathcal{K}$. Thus

$$q(x) \prod_{k\in\mathcal{K}} p(x;\theta_k) \leq q(x) \quad \forall x \in \mathcal{X} \setminus \mathcal{Y}_{\mathcal{K}}. \qquad (20)$$

Finally, we have that $q(x)$ and $\{p(x;\theta_k)\}_{k=1}^{K}$ are bounded probability densities. Thus there exists a finite $B > 0$ such that for each $x \in \mathcal{X}$ we have

$$\begin{aligned}
q(x) &\leq B, \qquad (21) \\
p(x;\theta_k) &\leq B. \qquad (22)
\end{aligned}$$

We now have the necessary ingredients to construct a finite upper bound on the integral in the statement of the theorem and thereby show that it is finite. We proceed as follows.

$$\begin{aligned}
\int_{\mathcal{X}} q(x) \prod_{k\in\mathcal{K}} p(x;\theta_k)\,dx &= \int_{\mathcal{Y}_{\mathcal{K}}} q(x) \prod_{k\in\mathcal{K}} p(x;\theta_k)\,dx + \int_{\mathcal{X}\setminus\mathcal{Y}_{\mathcal{K}}} q(x) \prod_{k\in\mathcal{K}} p(x;\theta_k)\,dx, \\
&\overset{(i)}{\leq} B^{1+|\mathcal{K}|} \int_{\mathcal{Y}_{\mathcal{K}}} dx + \int_{\mathcal{X}\setminus\mathcal{Y}_{\mathcal{K}}} q(x)dx, \\
&\overset{(ii)}{\leq} B^{1+|\mathcal{K}|} |\mathcal{Y}_{\mathcal{K}}| + \int_{\mathcal{X}} q(x)dx, \\
&\overset{(iii)}{\leq} (1 + |\mathcal{K}|) B^{1+|\mathcal{K}|} + 1, \\
&< \infty.
\end{aligned}$$

where, the inequality in step (i) follows from the upper bounds in (20), (21), and (22), the inequality in step (ii) follows since $\mathcal{X} \setminus \mathcal{Y}_{\mathcal{K}} \subseteq \mathcal{X}$, and the final inequality in step (iii) follows from the upper bound in (19).

Since the subset $\mathcal{K}$ that we fixed at the begining of the proof was an arbitrary one, the argument presented above will hold for any $\mathcal{K}$. Hence the integral in the statement of the theorem will be finite for all subsets $\mathcal{K}$ of $\{1, 2, \ldots, K\}$. This completes the proof. ■

Appendix B. Proof Of Theorem 6

In this appendix we prove Theorem 6 from Section 4. We begin with a basic result from real analysis that will play a key role in the proof.

Proposition 10 (Hölder's Inequality) *Let $\{f_k\}_{k=1}^K$ be a collection of measurable functions $f_k\colon \mathcal{X} \to \mathbb{R}$ on $\mathcal{X} \subseteq \mathbb{R}^d$. If $\hat{\alpha} = \{\alpha_k\}_{k=1}^K$ is such that each $\alpha_k \in [1,\infty]$ and $\sum_{k=1}^K \frac{1}{\alpha_k} = 1$, then we have*

$$\int_{\mathcal{X}} \prod_{k=1}^K |f_k(x)| dx \leq \prod_{k=1}^K \left(\int_{\mathcal{X}} |f_k(x)|^{\alpha_k} dx \right)^{\frac{1}{\alpha_k}},$$

where $|\cdot|$ denotes the absolute value and by standard convention, for $\alpha_k = \infty$ we define $\frac{1}{\alpha_k} = 0$ and

$$\left(\int_{\mathcal{X}} |f_k(x)|^{\alpha_k} dx \right)^{\frac{1}{\alpha_k}} = \operatorname*{ess\,sup}_{\mathcal{X}} |f_k|.$$

Proof See Wheeden and Zygmund (1977, Theorem 8.6) and use induction on K. ■

Next, we present a proof of Theorem 6. For the sake of readability, we repeat the statement of the theorem before presenting its proof.

Theorem *Let $q(x) \leq 1$ and $r(x) \leq 1$. If $\hat{a} = \{a_k\}_{k=1}^K$ is such that each $a_k \in [0,1]$ and $\sum_{k=1}^K a_k = 1$, then we have*

$$-\log\left(\int_{\mathcal{X}} q(x) \prod_{k=1}^K \left[e^{\langle \theta_k, t(x) \rangle} r(x) \right]^{z_k} dx \right) \geq -\sum_{k=1}^K z_k a_k \Upsilon\left(\frac{\theta_k}{a_k} \right),$$

where, for any $a_k = 0$ we define

$$a_k \Upsilon\left(\frac{\theta_k}{a_k} \right) = \log\left(\operatorname*{ess\,sup}_{\mathcal{X}} e^{\langle \theta_k, t(x) \rangle} r(x) \right).$$

Proof Let us fix an arbitrary z and let $\mathcal{K} \subseteq \{1, 2, \ldots, K\}$ denote which of its components are set to 1. Thus for this configuration of z, we will have $z_k = 1$ if $k \in \mathcal{K}$ and $z_k = 0$ if $k \notin \mathcal{K}$. We can see that there will be a unique $\mathcal{K}$ corresponding to each one of the 2^K possible configurations of z. Further, for a given $\hat{a}$ that is held fixed, let us define

$$a_q = 1 - \sum_{k \in \mathcal{K}} a_k.$$

From the definition of $\hat{a}$, we have that each $a_k \geq 0$ and since $\mathcal{K} \subseteq \{1, 2, \ldots, K\}$, we will additionally have

$$\sum_{k \in \mathcal{K}} a_k \leq \sum_{k=1}^K a_k = 1.$$

Combining these two observations with the definition of a_q, we can see that $a_q \in [0, 1]$. Thus, for the reciprocals of a_q and $\{a_k\}_{k\in\mathcal{K}}$, we will get

$$\frac{1}{a_q} \geq 1, \; \frac{1}{a_k} \geq 1, \tag{23}$$

$$a_q + \sum_{k\in\mathcal{K}} a_k = 1. \tag{24}$$

We now have a valid set of exponents for which Hölder's inequality is applicable. For the configuration of z that was fixed at the begining of the proof, we can start constructing a bound on the integral in the statement of the theorem as follows.

$$\begin{aligned}\int_{\mathcal{X}} q(x) \prod_{k=1}^{K} \left[e^{\langle\theta_k, t(x)\rangle} r(x)\right]^{z_k} dx &= \int_{\mathcal{X}} q(x) \prod_{k\in\mathcal{K}} e^{\langle\theta_k, t(x)\rangle} r(x) dx, \\ &\overset{(i)}{\leq} \left(\int_{\mathcal{X}} q(x)^{\frac{1}{a_q}} dx\right)^{a_q} \prod_{k\in\mathcal{K}} \left(\int_{\mathcal{X}} e^{\left\langle \frac{\theta_k}{a_k}, t(x)\right\rangle} r(x)^{\frac{1}{a_k}} dx\right)^{a_k}, \end{aligned} \tag{25}$$

where, the inequality in step (i) follows from (23), (24), and Hölder's inequality. Note that since our functions are probability densities, they will all be non-negative and so we can dispense with the absolute value signs. Just as in Hölder's inequality, by standard convention, for any $a_k = 0$ we define $\frac{1}{a_k} = \infty$ and

$$\left(\int_{\mathcal{X}} e^{\left\langle \frac{\theta_k}{a_k}, t(x)\right\rangle} r(x)^{\frac{1}{a_k}} dx\right)^{a_k} = \operatorname*{ess\,sup}_{\mathcal{X}} e^{\langle\theta_k, t(x)\rangle} r(x).$$

Let us now turn our attention to the leading term of (25). As $a_q \in [0, 1]$, there are two possibilities: $a_q \in (0, 1]$ and $a_q = 0$. First, suppose that $a_q \in (0, 1]$. Since $q(x) \leq 1$ and $\frac{1}{a_q} \geq 1$, we will have

$$\left(\int_{\mathcal{X}} q(x)^{\frac{1}{a_q}} dx\right)^{a_q} \leq \left(\int_{\mathcal{X}} q(x) dx\right)^{a_q} = 1.$$

Next, suppose that $a_q = 0$. Since $q(x) \leq 1$, we will have

$$\left(\int_{\mathcal{X}} q(x)^{\frac{1}{a_q}} dx\right)^{a_q} = \operatorname*{ess\,sup}_{\mathcal{X}} q(x) \leq 1.$$

Hence in either case, for $a_q \in [0, 1]$ we get

$$\left(\int_{\mathcal{X}} q(x)^{\frac{1}{a_q}} dx\right)^{a_q} \leq 1. \tag{26}$$

Moreover, since $r(x) \leq 1$ and each $\frac{1}{a_k} \geq 1$, we also get

$$r(x)^{\frac{1}{a_k}} \leq r(x). \tag{27}$$

Using the previous two results, we can get a further bound on the previous upper bound in (25) on the integral in the statement of the Theorem. We proceed as follows.

$$\begin{aligned}\Big(\int_{\mathcal{X}} q(x)^{\frac{1}{a_q}} dx\Big)^{a_q} \prod_{k\in\mathcal{K}}\Big(\int_{\mathcal{X}} e^{\left\langle \frac{\theta_k}{a_k}, t(x)\right\rangle} r(x)^{\frac{1}{a_k}} dx\Big)^{a_k} &\overset{(i)}{\leq} \prod_{k\in\mathcal{K}}\Big(\int_{\mathcal{X}} e^{\left\langle \frac{\theta_k}{a_k}, t(x)\right\rangle} r(x)^{\frac{1}{a_k}} dx\Big)^{a_k},\\ &\overset{(ii)}{\leq} \prod_{k\in\mathcal{K}}\Big(\int_{\mathcal{X}} e^{\left\langle \frac{\theta_k}{a_k}, t(x)\right\rangle} r(x) dx\Big)^{a_k}, \qquad (28)\end{aligned}$$

where, the inequality in step (i) follows from the upper bound in (26) and the inequality in step (ii) follows from the upper bound in (27).

Taking a log on the right-hand side of the the previous expression, multiplying it by -1, and using the definition of the log-partition function Υ given in (2), we get

$$\begin{aligned}-\log\Big(\prod_{k\in\mathcal{K}}\Big(\int_{\mathcal{X}} e^{\left\langle \frac{\theta_k}{a_k}, t(x)\right\rangle} r(x) dx\Big)^{a_k}\Big) &= -\sum_{k\in\mathcal{K}} a_k \Upsilon\Big(\frac{\theta_k}{a_k}\Big),\\ &\overset{(i)}{=} -\sum_{k\in\mathcal{K}} z_k a_k \Upsilon\Big(\frac{\theta_k}{a_k}\Big) - \sum_{k\notin\mathcal{K}} z_k a_k \Upsilon\Big(\frac{\theta_k}{a_k}\Big),\\ &= -\sum_{k=1}^{K} z_k a_k \Upsilon\Big(\frac{\theta_k}{a_k}\Big), \qquad (29)\end{aligned}$$

where, the equality in step (i) follows since $z_k = 1$ for $k \in \mathcal{K}$ and $z_k = 0$ for $k \notin \mathcal{K}$ according to the configuration of z that we fixed at the begining of the proof. If any $a_k = 0$, then we can simply define

$$a_k \Upsilon\Big(\frac{\theta_k}{a_k}\Big) = \log\Big(\operatorname*{ess\,sup}_{\mathcal{X}} e^{\langle \theta_k, t(x)\rangle} r(x)\Big).$$

Finally, combining the bounds in (25) and (28) together with the expression in (29) and using the fact that the $-\log(\cdot)$ function is monotone decreasing, we get the following lower bound.

$$-\log\Big(\int_{\mathcal{X}} q(x) \prod_{k=1}^{K} \big[e^{\langle \theta_k, t(x)\rangle} r(x)\big]^{z_k} dx\Big) \geq -\sum_{k=1}^{K} z_k a_k \Upsilon\Big(\frac{\theta_k}{a_k}\Big).$$

Note that the lower bound in the previous expression is not dependent on a particular $\mathcal{K}$ and hence a particular configuration of z. It will hold for any arbitrary $\mathcal{K} \subseteq \{1, 2, \ldots, K\}$ and hence any arbitrary configuration of z since there is a unique $\mathcal{K}$ corresponding to each possible configuration of z. Thus, the final lower bound will hold for all 2^K possible configurations of z. This completes the proof. ■

Appendix C. Proof Of Lemma 7

In this appendix we prove Lemma 7 from Section 4. For the sake of readability, we repeat the statement of the lemma before presenting its proof.

Lemma *Let* $h^*\colon \mathbb{R} \to \mathbb{R}_{++}$ *be the function* $h^*(y) = \log(1+e^y)$ *and* $h\colon [0,1] \to [-\log 2, 0]$ *be the negative binary entropy function* $h(b) = b\log b + (1-b)\log(1-b)$*. Then* h^* *and* h *are convex conjugates of one another and for* $y \in \mathbb{R}$*,* $b \in [0,1]$*, we will have*

$$h^*(y) = \max_{b\in[0,1]}\{by - h(b)\} \quad \textit{where} \quad \arg\max_{b\in[0,1]}\{by - h(b)\} = \frac{1}{1+e^{-y}},$$
$$h(b) = \max_{y\in\mathbb{R}}\{by - h^*(y)\} \quad \textit{where} \quad \arg\max_{y\in\mathbb{R}}\{by - h^*(y)\} = \log\left(\frac{b}{1-b}\right),$$

and

$$h^*(y) + h(b) \geq by.$$

Proof We can clearly see that h is a proper, continuously differentiable and thereby closed function. Also, we can verify via its second derivative, that h is a convex function as well. Its convex conjugate, or Legendre transform, is then defined by

$$\sup_{b\in[0,1]}\{by - h(b)\}.$$

This supremum will be finite and achievable for all $y \in \mathbb{R}$. Hence, the convex conjugate of h is defined for each $y \in \mathbb{R}$ and we can replace the sup with a max. Since h is a convex function, this maximum will be uniquely achieved. Setting the first derivative of the term within the sup in the previous expression to zero, we get

$$y - \log\left(\frac{b}{1-b}\right) = 0,$$

and solving this equation for b we get

$$b = \frac{1}{1+e^{-y}}.$$

We can see that this b lies in the interval $[0,1]$ and that the first derivative of the function to be maximized is zero at it. Combined with the convexity of h, we can conclude that

$$\arg\max_{b\in[0,1]}\{by - h(b)\} = \frac{1}{1+e^{-y}}.$$

Now, substituting the optimal value of b into the function to be maximized, we get

$$\begin{aligned}\max_{b\in[0,1]}\{by - h(b)\} &= \frac{y}{1+e^{-y}} - \frac{1}{1+e^{-y}}\log\left(\frac{1}{1+e^{-y}}\right) - \frac{e^{-y}}{1+e^{-y}}\log\left(\frac{e^{-y}}{1+e^{-y}}\right),\\ &= \frac{ye^y}{1+e^y} - \frac{ye^y}{1+e^y} + \frac{e^y}{1+e^y}\log(1+e^y) + \frac{1}{1+e^y}\log(1+e^y),\\ &= log(1+e^y),\\ &= h^*(y).\end{aligned}$$

We will proceed in exactly the same way for h^* as we did for h. It is a proper, continuously differentiable and thereby closed function. In addition, it is a convex function as well. The convex conjugate, or Legendre transform, of h^* is then defined by

$$\sup_{y \in \mathbb{R}} \{by - h^*(y)\}.$$

This supremum will be finite and achievable if $b \in [0, 1]$. Hence, the convex conjugate of h^* is defined for $b \in [0, 1]$ in which case we can replace the sup with a max. Since h^* is a convex function, this maximum will be uniquely achieved. Setting the first derivative of the term within the sup in the previous expression to zero, we get

$$b - \frac{1}{1 + e^{-y}} = 0,$$

and solving this equation for y we get

$$y = \log\left(\frac{b}{1 - b}\right).$$

We can see that this y is a well defined real number and that the first derivative of the function to be maximized is zero at it. Combined with the convexity of h^*, we can conclude that

$$\arg\max_{y \in \mathbb{R}} \{by - h^*(y)\} = \log\left(\frac{b}{1 - b}\right).$$

Now, substituting the optimal value of y into the function to be maximized, we get

$$\begin{aligned}\max_{y \in \mathbb{R}} \{by - h^*(y)\} &= b \log\left(\frac{b}{1 - b}\right) - \log\left(1 + \frac{b}{1 - b}\right), \\ &= b \log b - b \log(1 - b) - \log\left(\frac{1}{1 - b}\right), \\ &= b \log b + (1 - b) \log(1 - b), \\ &= h(b).\end{aligned}$$

Finally, from Fenchel's inequality for a proper convex function and its conjugate, we get

$$h^*(y) + h(b) \geq by.$$

This completes the proof. ■

References

S. Amari and H. Nagaoka. *Methods Of Information Geometry.* American Mathematical Society, 2001.

A. Banerjee, C. Krumpelman, S. Basu, R. Mooney, and J. Ghosh. Model based overlapping clustering. In *Proceedings Of The International Conference on Knowledge Discovery and Data Mining*, 2005a.

A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal Of Machine Learning Research*, 6:1705–1749, 2005b.

O. Barndorff-Nielsen. *Information And Exponential Families In Statistical Theory.* Wiley, 1978.

D. Bertsekas. *Nonlinear Programming.* Athena Scientific, 1999.

K. Heller and Z. Ghahramani. A nonparametric bayesian approach to modeling overlapping clusters. In *Proceedings Of The 11th International Conference On AI And Statistics*, 2007.

G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

J. Hiriart-Urruty and C. Lemaréchal. *Fundamentals Of Convex Analysis.* Springer, 2001.

M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

G. McLachlan and T. Krishnan. *The EM Algorithm And Extensions.* Wiley Interscience, 1996.

R. Redner and H. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984.

R. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, University of California at Berkeley, 2003.

R. Wheeden and A. Zygmund. *Measure And Integral: An Introduction To Real Analysis.* Marcel Dekker, 1977.

C. Williams, F. Agakov, and S. Felderhof. Products of gaussians. In *Advances In Neural Information Processing Systems 14*, 2002.