Clustering with Beta Divergences

Anonymous Author 1
Unknown Institution 1

Anonymous Author 2
Unknown Institution 2

Anonymous Author 3
Unknown Institution 3

Abstract

Clustering algorithms start with a fixed divergence metric, which captures the possibly asymmetric distance between two samples. In a mixture model, the sample distribution plays the role of a divergence metric. It is often the case that the distributional assumption is not validated, which calls for an adaptive approach. We consider a richer model where the underlying distribution belongs to a parametrized exponential family, called Tweedie Models. We first show the connection between the Tweedie models and beta divergences, and derive the corresponding hard-assignment clustering algorithm. We exploit this connection to identify moment conditions and use Generalized Method of Moments(GMoM) to learn the data distribution. Based on this adaptive approach, we propose four new hard clustering algorithms and compare them to the classical k-means and DP-means on synthetic data as well as seven UCI datasets and one large gene expression dataset. We further compare the GMoM routine to an approximate maximum likelihood routine and validate the computational benefits of the GMoM approach.

1 Introduction

Despite the general tendency towards more complex models and algorithms, k-means remains one of the most popular clustering algorithms [17]. In essence, k-means makes three assumptions: i) clusters are disjoint, implying a hard-assignment of samples to clusters, ii) the number of clusters is fixed and specified ahead of time, and iii) the distance between two samples is measured with Euclidian distance [3]. Many extensions to k-means have been proposed to relax these assumptions. The soft k-means algorithm allows for a soft-assignment of samples to clusters. In fact, it has been shown that k-means is the equivalent of EM for a Gaussian mixture model under the asymptotically small variance assumption [3]. This probabilistic model connection opened the door to using the DP-means algorithm, removing the second assumption of static numbers of clusters. DP-means was derived using the EM algorithm for a Dirichlet process (DP) Gaussian mixture model [17]. Finally, the (homoskedastic) Gaussian distribution implicit in the third assumption of Euclidean distance has been generalized to regular exponential family distributions [1]. Banerjee et al.'s bijection between regular exponential family distributions and regular Bregman divergences made it possible to write down the corresponding hard clustering algorithm for any given regular exponential family mixture model [1]. Furthermore, DP-means may be used with any regular Bregman divergence, producing a nonparametric hard clustering with any regular exponential family distribution [15].

After all these relaxations, we are still bound to specify a fixed divergence metric to measure the distance between two points. In the case of mixture models, this corresponds to choosing an underlying distribution. This is closely related to the model selection problem. In practice, the choice of distribution is difficult i) when we do not have a good idea about the natural distribution of the data; and ii) when it is not feasible to select the most appropriate exponential family distribution for each feature dimension. We attempt to solve this problem by enriching the model and proposing an adaptive algorithm. We say that the mixture model belongs to a family where each mixture component exhibits a certain characteristic, namely, that the sample variance is a scaled power function of the mean [16]. Mixture models with an exponential family distribution that captures this specific property of data are called Tweedie mixture models (TMM); TMMs are equally capable of representing Gaussian, Gamma, Compound Poisson-Gamma, Poisson, and inverse Gaussian mixture models [16].

Preliminary work. Under review by AISTATS 2016. Do not distribute.

In section 2, we start with the formal definitions of regular exponential family, regular Bregman divergences, Tweedie models, and beta divergences. We then state the known bijection between regular exponential family and regular Bregman divergences [1]. In the next section, we establish the connection between Tweedie models and beta divergences through the bijection. In doing this, we first show that the beta divergences are a subfamily of regular Bregman divergences. This is an extension of the results from [5, 14]. We then use the bijection and show that the corresponding regular Exponential family is the Tweedie models. Similar arguments have been made in [8, 24] with more restrictive definitions of beta divergences. The straightforward implication of this connection is that Bregman hard clustering and Bregman DP-Means algorithms can be modified to work with beta divergences, and they are the limiting cases of EM for TMMs and DP-TMMs, respectively. Section 4 explores this relationship.

In Section 5, we turn our attention to the estimation problem for the underlying distribution. We discuss the likelihood based and moment based approaches and show that Generalized Method of Moments (GMoM) is particularly suitable to the hard clustering problem. We then propose three new algorithms for the parametric hard clustering. ML-BHC and GMoM-BHC combines beta hard clustering(BHC) with the maximum likelihood routine and GMoM routine respectively. We also propose a purely moment based clustering algorithm(GMoM-HC). Finally we look at GMoM-BDPM, a nonparametric hard clustering algorithm that uses the GMoM routine with DP-Means. We compare the accuracy and runtime of these four algorithms with k-means and DP-Means on synthetic data, UCI datasets, and a large genomics dataset.

2 Background

We begin with the formal definitions of regular exponential family and regular Bregman divergences and state the bijection between them.

Definition 1. *A family of distributions* $\mathcal{F}_\Psi = \left\{p_{(\Psi,\Theta)} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} = int(\boldsymbol{\Theta}) = dom(\Psi) \subseteq \mathbb{R}^d\right\}$ *is called a* regular exponential family *if*

$$p_{(\Psi,\theta)}(\boldsymbol{x}) = \exp(\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle - \Psi(\boldsymbol{\theta}) - h(\boldsymbol{x})) \qquad (1)$$

where $\Psi(\boldsymbol{\theta})$ *is the log partition function,* $h(\boldsymbol{x})$ *is the base measure and the minimal sufficient statistics* $T(\boldsymbol{x}) = x$.

It is important to note that $\boldsymbol{\Theta}$ is an open set, and the pair $(\Psi, \boldsymbol{\Theta})$ is a Legendre type convex function [1]. Furthermore, we can express the expectation and variance as

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \triangledown\Psi(\boldsymbol{\theta})$$
$$\boldsymbol{\sigma^2}(\boldsymbol{\theta}) = \triangledown^2\Psi(\boldsymbol{\theta}).$$

Definition 2 (Bregman, 1967)**.** *Let* $\phi : S \to \mathbb{R}$ *be a strictly convex function defined on a convex set* $S \subseteq \mathbb{R}^d$ *such that* ϕ *is differentiable on nonempty* $ri(S)$*. The* Bregman divergence $D_\phi : S \times ri(S) \to \mathbb{R}_+$ *is defined as*

$$D_\phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \boldsymbol{x} - \boldsymbol{y}, \triangledown\phi(\boldsymbol{y})\rangle. \qquad (2)$$

where $\triangledown\phi(\boldsymbol{y})$ *is the gradient vector of* ϕ *evaluated at* $\boldsymbol{y}$.

Definition 3 (Banerjee et al., 2005)**.** *Let* $f : \boldsymbol{\Theta} \to \mathbb{R}_{++}$ *be continuous exponentially convex function such that* $\boldsymbol{\Theta}$ *is open and* $\Psi(\boldsymbol{\theta}) = \log(f(\boldsymbol{\theta}))$ *is strictly convex. Let* ϕ *be the conjugate function of* Ψ*. Then the Bregman divergence* D_ϕ *derived from* ϕ *is called a* regular Bregman divergence.

Regular Bregman divergences are a subset of Bregman divergences. The main difference comes in the openness of the domain for the conjugate of the generating convex function.

Theorem 1 (Banerjee et al., 2005)**.** *There is a bijection between the regular exponential family and the regular Bregman divergences.*

The bijection implies that Eq. 1 can be written as

$$p_{(\Psi,\theta)}(\boldsymbol{x}) = p_{(\phi,\boldsymbol{\mu})}(\boldsymbol{x}) = \exp(-d_\phi(\boldsymbol{x}, \boldsymbol{\mu}))f_\phi(\boldsymbol{x}), \qquad (3)$$

where ϕ is the convex conjugate of Ψ and $f_\phi(\boldsymbol{x}) = \exp(\phi(\boldsymbol{x}) - h(\boldsymbol{x}))$. Using the bijection, [1] shows that k-means algorithm can be generalized to work with any regular Bregman divergence. The resulting algorithm is called Bregman hard clustering. Furthermore, we know that Bregman hard clustering is the limiting case of EM for regular exponential family mixture models under the asymptotically small variance assumption [1].

Next we look at a parametrized family of divergences called beta divergences. Beta divergences were first introduced for $\beta > 1$ [2] and later generalized to $\beta \in \mathbb{R}$ [7]. There is no consensus on the domain choice for the beta divergences [2, 20, 8, 21, 5, 14, 6, 23]. We will define it to be as inclusive as possible while preserving the desired convexity propery. The formal definition is given by

Definition 4. *Let* S *be* $\mathbb{R}_{++}$ *for* $\beta < 0$*,* $\mathbb{R}_+$ *for* $\beta \geq 0$ *and* $\mathbb{R}$ *for* $\beta = 2$*, then beta divergence* $D_\beta : S \times ri(S) \to$

$\mathbb{R}_+$ *is defined as*

$$D_\beta(x,y) \doteq \begin{cases} \frac{x^\beta + (\beta-1)y^\beta - \beta x y^{\beta-1}}{\beta(\beta-1)} & \beta \in \mathbb{R} \setminus \{0,1\} \\ x(\log x - \log y) + (y-x) & \beta = 1 \\ \frac{x}{y} - \log\left(\frac{x}{y}\right) - 1 & \beta = 0. \end{cases} \tag{4}$$

Together with the limiting cases, this definition of beta divergence smoothly connects the Itakura-Saito divergence ($\beta = 0$) to Euclidean distance ($\beta = 2$) while passing through KL divergence ($\beta = 1$). Beta divergence has been used in the context of principal component analysis (PCA) [21], independent component analysis (ICA) [20], and nonnegative matrix factorization (NMF) [6].

Finally, we look at Tweedie models, a subfamily of the regular exponential family. The Tweedie models include Gaussian, Gamma, Compound Poisson-Gamma, Poisson, Inverse Gaussian, and positive stable distributions [16].

Definition 5 (Jorgensen, 1987)**.** *Let $\kappa \in \mathbb{R}_{++}$ and $\rho \in \mathbb{R}$ be shape and dispersion parameters, respectively. An exponential dispersion model satisfying*

$$\sigma^2(\theta) = \kappa \mu(\theta)^\rho \tag{5}$$

is called a Tweedie model.

In general, Tweedie models do not have a closed form density functions. Two possible approximations to the density function are proposed using an infinite summation [9] and using Fourier inversion [10].

3 Tweedie Models and Beta Divergences

In this section, we explore the connection between beta divergences and Tweedie models. Specifically, we show that i) beta divergences are a subfamily of regular Bregman divergences, and ii) the corresponding regular exponential family is the Tweedie models. Previously, [5, 14] showed that beta divergences are a subfamily of Bregman divergences. We broaden the definition of beta divergences(Def. 4) and show that they are a subfamily of the more restrictive regular Bregman divergences. Furthermore [8, 24] noted that the density of a Tweedie model can be written in terms of a beta divergence. We show that this results still holds for the proposed beta divergences. More importantly, we provide a careful treatment of the boundaries. After making the connection between beta divergences and Tweedie models using the bijection in Theorem 1. In Section 4, we will use (i) to modify Bregman hard clustering algorithms to work with beta divergences and use (ii) to assert that the resulting algorithm is the limiting case of EM for TMM. In Section 5, we will use (ii) to write down moment conditions for Beta hard clustering.

Remark 1. *The density of a Tweedie model characterized by $\sigma^2(\theta) = \kappa\mu(\theta)^{2-\beta}$ can be written in terms of beta divergence as*

$$p(x \mid \mu, \beta, \kappa) = \exp(-\frac{1}{\kappa} D_\beta(x, \mu)) f_{\beta,\kappa}(x) \tag{6}$$

where $f_{\beta,\kappa}(x)$ is the base measure.

Proof. Let $\psi_\beta(\theta) : \Theta \to \mathbb{R}$ be defined

$$\psi_\beta(\theta) \doteq \begin{cases} \frac{1}{\beta}(((\beta-1)\theta + 1)^{\frac{\beta}{\beta-1}} - 1) & \beta \in \mathbb{R} \setminus \{0,1\} \\ e^\theta - 1 & \beta = 1 \\ -\log(1-\theta) & \beta = 0 \end{cases} \tag{7}$$

with Θ given by

$$\Theta = \begin{cases} (-\infty, \frac{1}{1-\beta}) & \beta \in (-\infty, 1) \\ (\frac{1}{1-\beta}, \infty) & \beta \in (1,\infty) \setminus \{2\} \\ \mathbb{R} & \beta \in \{1, 2\}. \end{cases} \tag{8}$$

The first two derivatives are given by

$$\psi'_\beta(\theta) = \begin{cases} ((\beta-1)\theta + 1)^{\frac{1}{\beta-1}} & \beta \neq 1 \\ e^\theta & \beta = 1 \end{cases}$$

and

$$\psi''_\beta(\theta) = \begin{cases} ((\beta-1)\theta + 1)^{\frac{2-\beta}{\beta-1}} & \beta \neq 1 \\ e^\theta & \beta = 1 \end{cases}$$

Since $\psi''_\beta(\theta) > 0$ for every $\theta \in \Theta$, we conclude that ψ_β is strictly convex. Therefore, with the open set Θ, ψ_β defines a regular exponential family distribution as well as a regular Bregman divergence. Furthermore, we note that $\psi''_\beta(\theta) = \psi'_\beta(\theta)^{2-\beta}$, which indicates that this is a Tweedie model with shape parameter $2-\beta$ and dispersion parameter 1. This construction differs from an earlier definition of Tweedie model [16] in that we explicitly restrict the domain of the natural parameter to be an open set. In our construction, there is no degenerate distribution for values $1 < \beta < 2$. More recent work follows a similar path [10].

The convex conjugate of ψ_β is given by

$$\phi_\beta(t) = \begin{cases} \frac{1}{\beta(\beta-1)}(t^\beta - \beta t + \beta - 1) & \beta \in \mathbb{R} \setminus \{0,1\} \\ t\log t - t + 1 & \beta = 1 \\ t - \log t - 1 & \beta = 0 \end{cases} \tag{9}$$

Algorithm 1 Beta Hard Clustering (BHC)

Input: Data $\mathcal{X} = \{\boldsymbol{x_i}\}_{i=1}^N$, shape vector $\boldsymbol{\beta}$, number of clusters k.
Output: Partition $\mathcal{C}(\mathcal{X}) = \{\mathcal{X}_h\}_{h=1}^k$
Initialize $\{\boldsymbol{\mu_h}\}_{h=1}^k$.
repeat
  $\boldsymbol{\mathcal{X}_h} \leftarrow \varnothing$ for $h = 1, \ldots, k$
  for $i = 1$ **to** N **do**
    $h = \underset{h'}{\operatorname{argmin}}\, D_{\boldsymbol{\beta}}(\boldsymbol{x_i}, \boldsymbol{\mu_{h'}})$
    $\boldsymbol{\mathcal{X}_h} \leftarrow \boldsymbol{\mathcal{X}_h} \cup \{\boldsymbol{x_i}\}$
  end for
  for $h = 1$ **to** k **do**
    $\boldsymbol{\mu_h} \leftarrow \frac{1}{|\boldsymbol{\mathcal{X}_h}|} \sum_{\boldsymbol{x_i} \in \boldsymbol{\mathcal{X}_h}} \boldsymbol{x_i}$
  end for
until convergence

with the domain $dom(\phi) = S$ as given in Def. 4. The regular Bregman divergence generated by ϕ_β is the beta divergence defined in Eq. 4.

For a Tweedie model with an arbitrary dispersion parameter κ, we need to scale the natural parameter, θ. Using Lemma 3.1 of [15], we write

$$\begin{aligned}
\tilde{\theta} &= \theta/\kappa \\
\tilde{\Psi}(\tilde{\theta}) &= \Psi(\kappa\tilde{\theta})/\kappa \\
\tilde{\phi}(\tilde{t}) &= \phi(\tilde{t})/\kappa \\
D_{\tilde{\phi}}(x, y) &= D_{\phi}(x, y)/\kappa \\
\nabla\tilde{\Psi}(\tilde{\theta}) &= \nabla\Psi(\kappa\tilde{\theta}) \\
\nabla^2\tilde{\Psi}(\tilde{\theta}) &= \kappa\nabla^2\Psi(\kappa\tilde{\theta}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sigma^2(\tilde{\theta}) &= \nabla^2\tilde{\Psi}(\tilde{\theta}) = \kappa\nabla^2\Psi(\theta) = \kappa\nabla\Psi(\theta)^{2-\beta} \\
&= \kappa\nabla\tilde{\Psi}(\tilde{\theta})^{2-\beta} = \kappa\mu(\tilde{\theta})^{2-\beta}
\end{aligned}$$

which fully characterizes a Tweedie model. The rest follows from the bijection in Eq. 3. □

4 Tweedie Mixture Models and Beta Hard Clustering

We now consider a multidimensional Tweedie mixture model. Given the set of parameters $\boldsymbol{\lambda} = \langle \mu_{1j} \ldots \mu_{kd}, \beta_1 \ldots \beta_d, \kappa_1 \ldots \kappa_d \rangle$ and mixture weights $\boldsymbol{\pi} = \langle \pi_1 \ldots \pi_k \rangle$, the likelihood of $\mathcal{X} = \{\boldsymbol{x_i}\}_{i=1}^N$, $\mathcal{L}_{TMM}(\mathcal{X} \mid \boldsymbol{\lambda})$, is given by

$$\sum_{i=1}^{N} \sum_{h=1}^{k} \pi_h \sum_{j=1}^{d} \exp(-\frac{1}{\kappa_j} D_{\beta_j}(x_{ij}, \mu_{hj})) f_{\beta_j, \kappa_j}(x_{ij}). \tag{10}$$

Algorithm 2 Beta DP-Means (BDPM)

Input: Data $\mathcal{X} = \{\boldsymbol{x_i}\}_{i=1}^N$, shape vector $\boldsymbol{\beta}$, threshold ζ.
Output: Partition $\mathcal{C}(\mathcal{X}) = \{\mathcal{X}_h\}_{h=1}^k$
Initialize $\boldsymbol{\mu_1} = \frac{1}{N} \sum_{\boldsymbol{x_i}} \boldsymbol{x_i}$, $k = 1$.
repeat
  $\boldsymbol{\mathcal{X}_h} \leftarrow \varnothing$ for $h = 1, \ldots, k$
  for $i = 1$ **to** N **do**
    $h = \underset{h'}{\operatorname{argmin}}\, D_{\boldsymbol{\beta}}(\boldsymbol{x_i}, \boldsymbol{\mu_{h'}})$
    if $D_{\boldsymbol{\beta}}(\boldsymbol{x_i}, \boldsymbol{\mu_h}) \leq \zeta$ **then**
      $\boldsymbol{\mathcal{X}_h} \leftarrow \boldsymbol{\mathcal{X}_h} \cup \{\boldsymbol{x_i}\}$
    else
      $k \leftarrow k + 1$
      $\boldsymbol{\mathcal{X}_k} \leftarrow \{\boldsymbol{x_i}\}$
    end if
  end for
  for $h = 1$ **to** k **do**
    $\boldsymbol{\mu_h} \leftarrow \frac{1}{|\boldsymbol{\mathcal{X}_h}|} \sum_{\boldsymbol{x_i} \in \boldsymbol{\mathcal{X}_h}} \boldsymbol{x_i}$
  end for
until convergence

We used the linearity of Bregman divergences to allow each dimension to have a separate shape and dispersion parameter. This gives us the flexibility to adapt to the most appropriate distribution along each dimension.

Under the asymptotically small variance assumption, the EM algorithm for a regular exponential family mixture model reduces to Bregman hard clustering [1]. Following Remark 1, we conclude that as $\boldsymbol{\kappa} \to 0$, EM for TMM is equivalent to the Beta hard clustering algorithm (BHC) with the objective:

$$\mathcal{L}_{BHC}(\mathcal{C}(\mathcal{X}) \mid \boldsymbol{\lambda}) = \sum_{h=1}^{k} \sum_{\boldsymbol{x_i} \in \boldsymbol{\mathcal{X}_h}} \sum_{j=1}^{d} D_{\beta_j}(x_{ij}, \mu_{hj}), \tag{11}$$

where $\mathcal{C}(\mathcal{X}) = \{\mathcal{X}_h\}_{h=1}^k$ is the partition.

Similarly, we say that EM for Dirichlet Process Tweedie mixture model reduces to Beta DP-Means($BDPM$) with the objective $\mathcal{L}_{BHC}(\mathcal{C}(\mathcal{X}) \mid \boldsymbol{\lambda}) + \zeta k$ where ζ is the threshold parameter in DP-means [15]. It should be noted that both BHC and $BDPM$ require the shape parameter $\boldsymbol{\beta}$ to be specified up front.

5 Underlying Distribution

To identify the underlying data distribution, we investigate different approaches for estimating the shape vector $\boldsymbol{\beta}$. In the context of clustering, the most straightforward approach is to use the estimator $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}}\, \mathcal{L}_{BHC}$. Unfortunately, this only gives a triv-

ial solution [9]. In the context of mixture models, maximum likelihood estimate $\hat{\boldsymbol{\lambda}} = \text{argmin}_{\boldsymbol{\lambda}}\ \mathcal{L}_{TMM}$ can be used. The MLE estimates for the cluster centers are straigthforward and independent of $\boldsymbol{\beta}$ and $\boldsymbol{\kappa}$. Estimating the shape and dispersion vectors is however cumbersome. For generalized linear models(GLM) with Tweedie error distributions, it is possible to approximate the derivative of the likelihood with respect to $\boldsymbol{\kappa}$ and use grid search for $\boldsymbol{\beta}$ to maximize the approximated likelihood [10]. Clustering problem can easily be cast as a GLM and $\boldsymbol{\lambda}$ can be estimated. However, there are few issues with this approach. It is computationally costly, the resolution of the grid search is arbitrary and more importantly the approximations of the likelihood and its derivative require the parameter space to be partitioned. Therefore it is not feasible to search all possible shape and dispersion parameters in a single routine.

An alternative to the likelihood approach is the method of moments. For a sample set from a Tweedie distribution, we can estimate μ, β and κ using the first three moments. However, clustering problem exhibits a useful characteristics. The parameters $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$ are shared across clusters. Therefore the first two population moments for k clusters allow us to write down $2kd$ equations for $(k+2)d$ parameters. In other words, it is possible overidentify the system only with the first two population moments. This is the ideal setting for Generalized Method of Moments (GMoM) [12].

Based on the connection between Tweedie models and beta divergences established in Remark 1, we write the first two moments for the h^{th} cluster along dimension j as

$$\boldsymbol{m}_{hj}(\boldsymbol{x};\boldsymbol{\lambda}) = \left[x_j - \mu_{hj}\ ,\ x_j^2 - \mu_{hj}^2 - \kappa_j \mu_{hj}^{2-\beta_j}\right].$$

From Eq. 5, we know that $E[\boldsymbol{m}_{hj}(\boldsymbol{x};\boldsymbol{\lambda})] = \mathbf{0}$ for each cluster $h = \{1,\dots,k\}$ along each dimension $j = \{1,\dots,d\}$.

In Hansen's iterative GMoM method, the weight matrix is constructed by taking the inverse of the residual matrix [12]. This construction yields asymptotically efficient estimates for the parameter. In the hard clustering problem, clusters are disjoint; hence, the residual matrix and the weight matrix are block diagonal. The inversion of the weight matrix is then $O(kd)$ rather than $O(k^3 d)$. A variant of the iterative GMoM is the Continuously Updating Generalized Method of Moments (CUGMoM) [13]. We work with CUGMoM as it is computationally more reliable. The CUGMoM objective is given by

$$\mathcal{L}_{GMoM}(\mathcal{C}(\mathcal{X});\boldsymbol{\lambda}) = \sum_{h=1}^{k}\sum_{j=1}^{d} \bar{\boldsymbol{m}}_{\boldsymbol{hj}}{}^T \boldsymbol{W}_{hj}^{-1} \bar{\boldsymbol{m}}_{\boldsymbol{hj}} \quad (12)$$

where

$$\bar{\boldsymbol{m}}_{\boldsymbol{hj}} = \frac{1}{|\boldsymbol{\mathcal{X}_h}|}\sum_{\boldsymbol{x_i}\in\boldsymbol{\mathcal{X}_h}} \boldsymbol{m_{hj}}(\boldsymbol{x_i};\boldsymbol{\lambda}) \quad (13)$$

$$\boldsymbol{W}_{hj} = \frac{1}{|\boldsymbol{\mathcal{X}_h}|}\sum_{\boldsymbol{x_i}\in\boldsymbol{\mathcal{X}_h}} \boldsymbol{m_{hj}}(\boldsymbol{x_i};\boldsymbol{\lambda})\ \boldsymbol{m_{hj}}(\boldsymbol{x_i};\boldsymbol{\lambda})^T. \quad (14)$$

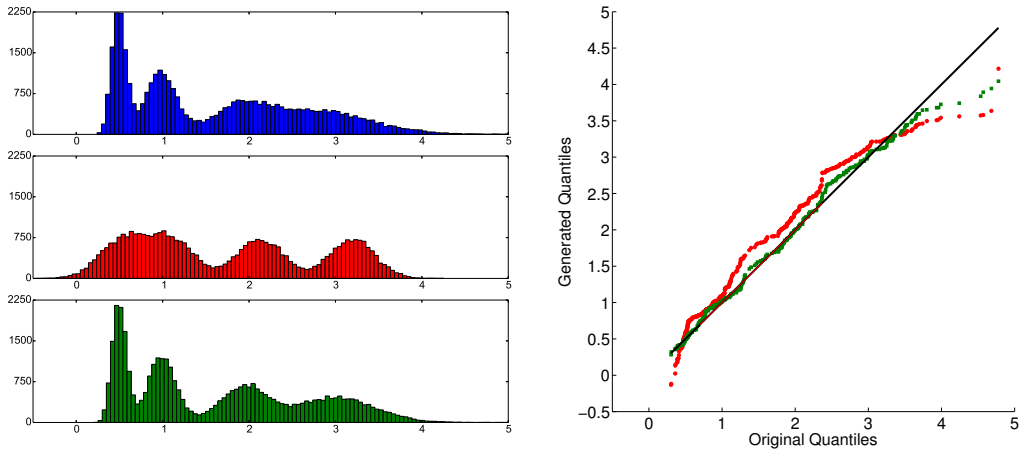
The CUGMoM estimate is $\hat{\boldsymbol{\lambda}} = \text{argmin}_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}}\ \mathcal{L}_{GMoM}(\mathcal{C}(\mathcal{X});\boldsymbol{\lambda})$. We note that the summary statistics required to compute the objective as well as its first derivative are the empirical raw moments up to fourth degree. To solve this optimization problem we used the L-BFGS-B algorithm [4]. The boundary conditions are determined by the maximally feasible region given data. As for the initialization, we first estimate β_j and κ_j within each cluster using the first three empirical raw moments. We then choose the median value across clusters along each dimension. Together with the empirical first population moments (MLE estimates of μ_{hj}), these form the initial value of $\boldsymbol{\lambda}$.

Algorithm 3 GMoM Hard Clustering

Input: Data $\{\boldsymbol{x_i}\}_{i=1}^N$, number of clusters k
Initialize $\boldsymbol{\lambda}$ and $\boldsymbol{W}_{hj} = I$
repeat
  $\boldsymbol{\mathcal{X}_h} \leftarrow \varnothing$ for $h = 1,\dots,k$
  for $i = 1$ **to** N **do**
    $h = \underset{h'}{\text{argmin}} \sum_{j=1}^{d} \boldsymbol{m}_{h'j}(\boldsymbol{x}_i)^T \boldsymbol{W}_{h'j}^{-1} \boldsymbol{m}_{h'j}(\boldsymbol{x}_i)$
    $\boldsymbol{\mathcal{X}_h} \leftarrow \boldsymbol{\mathcal{X}_h} \cup \{\boldsymbol{x_i}\}$
  end for
  $\boldsymbol{\lambda} = \text{argmin}_{\boldsymbol{\lambda'}\in\boldsymbol{\Lambda}}\ \mathcal{L}_{GMoM}(\mathcal{C}(\mathcal{X});\boldsymbol{\lambda'})$
  $\boldsymbol{W}_{hj} \leftarrow \boldsymbol{W}_{hj}(\mathcal{C}(\mathcal{X});\boldsymbol{\lambda})$
until convergence

Having figured out a way to estimate $\boldsymbol{\lambda}$ given $\mathcal{C}(\mathcal{X})$, we can combine GMoM with *BHC*. We iterate between CUGMoM estimates and *BHC* until the stopping criterion is met. We call the resulting algorithm *GMoM-BHC*. Similarly, we call the DP-Means variant *GMoM-BDPM*. Although there is no guarantee of convergence, one can use $\mathcal{L}_{GMoM}$, $\mathcal{L}_{BHC}$ or the cluster assignments as the stopping criterion.

A more direct approach is to use $\mathcal{L}_{GMoM}$ in both the assignment and maximization steps. The resulting algorithm is called the GMoM Hard Clustering (*GMoM-HC*) and has the same convergence properties as k-means. One can see this as a variant of k-means for the pseudo model characterized by $\mathcal{L}_{GMoM}$ [11].



(a) (b)

Figure 1: **Simulated density estimation results.** a) Histograms for the synthetic data of a gamma mixture model (blue), generated data from the Gaussian mixture model fit by k-means (red), generated data from the TMM fit by *GMoM-BHC* (green) b) Corresponding quantile-quantile plot for the empirical quantiles.

6 Results

6.1 Setup

We use Normalized Mutual Information (NMI) to quantify the match of the outcome of a clustering algorithm with ground truth. NMI produces values between 0 and 1, where 0 corresponds to random cluster assignments. For synthetic data experiments, we computed the empirical quantile-quantile plot using the samples from the posterior model.

We run each algorithm with random restarts. For the variants of k-means(*BHC*, *ML-BHC*, *GMoM-BHC* and *GMoM-HC*), centroids are initialized randomly. In DP-means variants (*BDPM* and *GMoM-BDPM*), we shuffle the data before each run.

To set the threshold parameter in DP-Means, *BDPM* ,and *GMoM-BDPM*, we use the farthest-first heuristic [17]. We start with a set Υ containing only the global mean. We iteratively add $\text{argmax}_{\boldsymbol{x}\in\mathcal{X}\setminus\Upsilon}\{\min_{\boldsymbol{s}\in\Upsilon} D_{\boldsymbol{\beta}}(\boldsymbol{x},\boldsymbol{s})\}$ to the set until the set size is $k+1$ for a given value k. We then set the threshold ζ to be $\min_{\boldsymbol{s}\in\Upsilon} D_{\boldsymbol{\beta}}(\tilde{\boldsymbol{x}},\boldsymbol{s})$ where $\tilde{\boldsymbol{x}}$ is the last element added to the set.

For the ML routine, we utilize *tweedie* package in R with Fourier inversion method [10]. We set the resolution of the grid search for the shape parameter to 0.25. The minimum value is -3.0 and the maximum value is 1.0 along each dimension.

As noted before, there is no guarantee of convergence for *GMoM-BHC* and *GMoM-BDPM*. We run each algorithm for 100 rounds, and select the model with the lowest inertia ($\mathcal{L}_{BHC}$). We terminate the algorithm early if the cluster assignments converge first. For a fair comparison, we use the same criterion for *ML-BHC*. In *BHC*, *BDPM* and *GMoM-HC*, we use the cluster assignments as the convergence criterion.

6.2 Synthetic Data Experiments

To illustrate the importance of distributional assumptions, we generate a simple single dimension synthetic dataset with four Gamma mixture components. Each cluster has 100 samples and the centroids are located at $0.5, 1.0, 2.0$ and 3.0. $\kappa = 0.03$ is fixed across clusters. This is a Tweedie mixture model (TMM) with $\beta = 0$. We first run k-means and sample from the corresponding posterior Gaussian mixture model. The parameters$(\mu_1, \mu_2, \mu_3, \mu_4, \Sigma)$ of the posterior model are set to be the MLE estimates given clusters. We then run *GMoM-BHC* and generate samples from the resulting TMM using [9]'s *tweedie* package. We see in the results that points generated from k-means deviate from the empirical quantiles significantly (Figure 1). A closer look reveals that both *GMoM-BHC* and k-means identify centroids quite accurately; however, *GMoM-BHC*'s β estimate (0.238) is closer to the true β value while k-means uses the fixed $\hat{\beta} = 2$.

Next, we compare *ML-BHC*,*GMoM-BHC* and *GMoM-HC* with k-means on single dimension synthetic datasets with 4, 8, and 16 clusters with fixed cluster size of 100. We look at Inverse Gaussian, Continuous Poisson, Compound Poisson Gamma, Gamma, and Gaussian mixture models with centroids at $\langle 2^{-k/2}, \ldots, 2^{k/2}\rangle$. Here $\kappa = 0.030$ except for Inverse

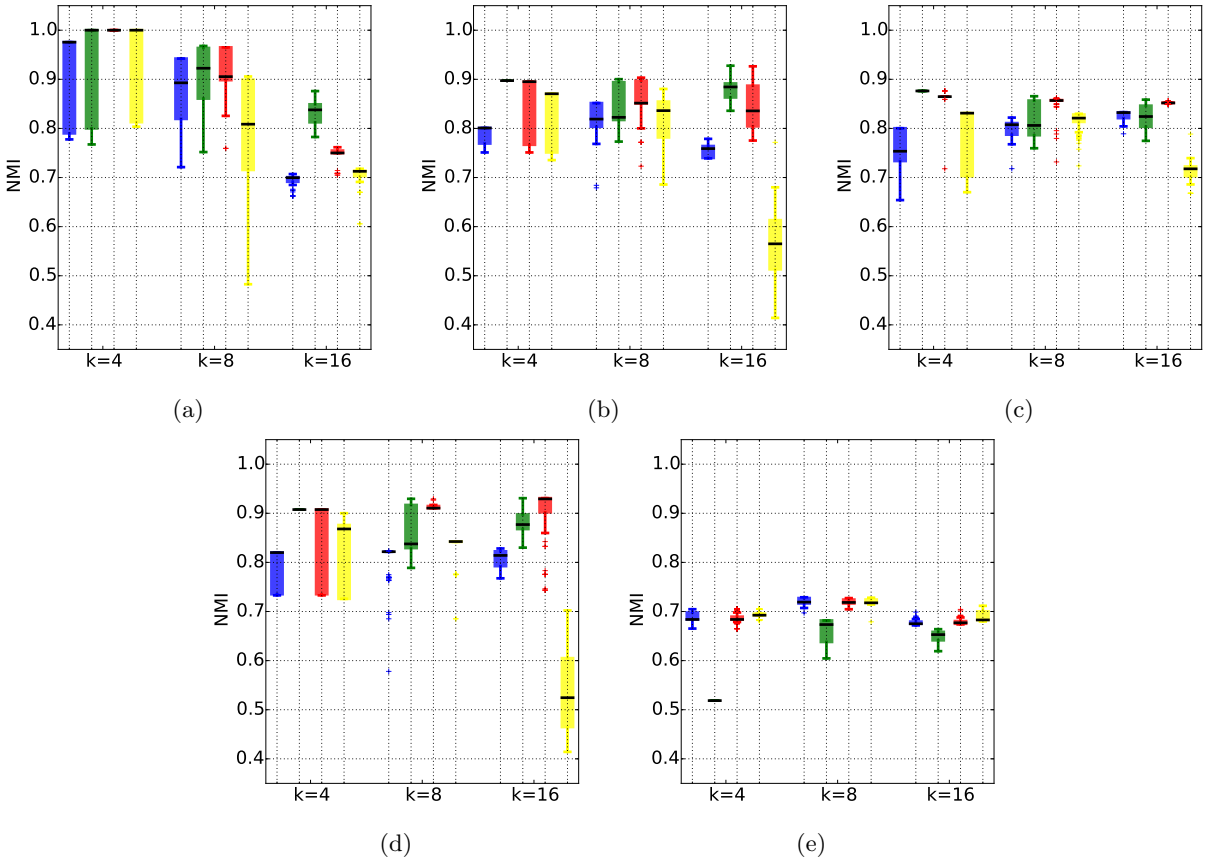


Figure 2: **Simulated clusters results.** Comparison of NMI value for k-means(blue), *ML-BHC*(green), *GMoM-BHC*(red) and *GMoM-HC*(yellow) on synthetic datasets with 4, 8, and 16 clusters and underlying distributions of a) Inverse Gaussian ($\beta = -1$), b) Continuous Poisson ($\beta = 0$), c) Compound Poisson Gamma ($\beta = 0.5$), d) Gamma ($\beta = 1$), and e) Gaussian ($\beta = 2$)

Gaussian where we use 0.008. We run each experiment 100 times and show the Whisker-Box plots in Figure 2. In the Gaussian case ($\beta = 2.0$), *GMoM-BHC* and *GMoM-HC* mimic k-means while *ML-BHC* underperforms slightly. This is due to the restrictions on the grid search. As we diverge from the Gaussian case (as β decreases), the relative performance of k-means decreases. Between *ML-BHC* and *GMoM-BHC*, the latter seems better except for the Inverse Gaussian case. This is again related to the grid search restrictions. *GMoM-BHC* outerperforms *GMoM-HC* as well. One likely explanation is that the underlying model of *GMoM-BHC* is closely related to the mixture model we use to generate data whereas pseudo-model of *GMoM-HC* does not necessarily represent a TMM.

6.3 UCI Data Experiments

So far we have looked at synthetic datasets with the same underlying distribution along each feature dimension. However, in real life this is not necessarily the case, and our proposed algorithms are capable of recognizing different distributional characteristics. We look at seven datasets from UCI repository [19]. We compare *ML-BHC*, *GMoM-BHC*, *GMoM-BDPM* and *GMoM-HC* with k-means and DP-means. We run each algorithm for 100 times and report the NMI value and the runtime of the model fit with the lowest inertia ($\mathcal{L}_{GMoM}$ in the case of *GMoM-HC*). Table1 and Table2 summarize the results. Based on pairwise comparisons, *GMoM-BHC* seems to be the best performing algorithm. The *GMoM-BHC* performances on *wine* and *steel* are particularly noteworthy. When we analyze the resulting TMM, we observe *Gamma* and *Poisson* characteristics along several dimensions. *ML-BHC* shows a comparable performance to *GMoM-BHC* except on *yeast* and *pima*. These are datasets with Gaussian characteristics and k-means performs relatively well in both. It is important to emphasize that the restrictions on the ML routine prevents *ML-BHC* to capture purely Gaussian characteristics. *GMoM-HC*, another adaptive algorithm we proposed, shows a comparable performance to *GMoM-BHC*. Between DP-Means and *GMoM-BDPM*, the latter seems

Table 1: Comparison of NMI values for k-means, DP-means and proposed algorithms on UCI datasets.

DATASET	N	D	K	KM	DPM	*ML-BHC*	*GMoM-BHC*	*GMoM-BDPM*	*GMoM-HC*
IRIS	149	4	2	0.869	0.869	**1.000**	0.944	0.732	0.838
PIMA	767	8	2	0.030	0.020	0.002	0.030	0.013	**0.114**
WINE	177	13	3	0.426	0.395	0.756	**0.769**	0.471	0.767
WINE Q.(R)	1598	11	6	0.043	0.038	0.059	**0.070**	0.051	0.049
WINE Q.(W)	4897	11	7	0.028	0.032	0.032	0.025	0.038	**0.056**
STEEL	1940	20	7	0.114	0.113	0.209	**0.238**	0.163	0.202
YEAST	1483	6	10	**0.261**	0.255	0.148	0.257	0.178	0.171

Table 2: Comparison of runtime in seconds for k-means, DP-means and proposed algorithms on UCI datasets.

DATASET	N	D	K	KM	DPM	*ML-BHC*	*GMoM-BHC*	*GMoM-BDPM*	*GMoM-HC*
IRIS	149	4	2	0.021	0.053	103.647	1.841	0.989	1.524
PIMA	767	8	2	0.406	0.627	148.539	37.302	1.767	11.455
WINE	177	13	3	0.069	0.123	611.870	6.587	0.505	1.568
WINE Q.(R)	1598	11	6	6.912	11.474	5493.993	315.089	977.549	30.811
WINE Q.(W)	4897	11	7	14.814	93.611	20473.482	1630.806	13501.052	222.957
STEEL	1940	20	7	8.279	8.864	2465.019	426.102	40.520	87.238
YEAST	1483	6	10	4.591	14.101	327.298	20.180	421.056	28.400

to show a better performance. As for the runtime, we observe that *GMoM-BHC* is an order of magnitude faster than *ML-BHC*. We also note that the runtime of k-means and *BHC* are similar and *GMoM-BHC* is at most 100 times slower than k-means as expected. Overall, we conclude that *GMoM-BHC* provides a good hybrid approach to adaptive clustering.

6.4 Genetics Data Experiments

We next looked at an RNA-sequencing gene expression dataset with $19K$ genes for 462 samples [18]. This technology takes a single sample of RNA, breaks it up into small pieces, sequences each of those pieces (*reads*), and maps the reads back to the reference human genome. The output is then a set of (smoothed) counts representing the number of reads that mapped to each gene; these data are often modeled with a negative binomial, overdispersed Poisson, or zero-inflated Poisson distribution. For these data, we establish "ground truth" with functional annotation clustering using three different databases (KEGG, REACTOME and GO) and the open source GeneSCF package [22].

We run k-means and *GMoM-BHC* 100 times with random initializations. As a common practice, we centered the data before running k-means to improve its performance. *GMoM-BHC* does not require this adjustment. Table3 gives the mean NMI values and the standard deviation for both algorithms with respect to different ontology databases. *GMoM-BHC* outperforms k-means in each case with a significant margin. When we analyzed the resulting TMM, we see that it has Compound Poisson-Gamma characteristic. This is a Gamma-like distribution with a point mass at zero. For gene expression, we believe this is a better model than Gaussian centered around a non-zero mean or the overdispersed Poisson, since it captures genes that are unexpressed in specific samples while preserving the statistical Gamma shape when expressed. The outcome of our experiment confirms this hypothesis.

Table 3: Comparison of NMI value for K-Means and *GMoM-BHC* on gene expression dataset

	KM	GMM-BHC
KEGG	0.013 ± 0.001	$\mathbf{0.022 \pm 0.001}$
REACTOME	0.018 ± 0.002	$\mathbf{0.028 \pm 0.002}$
GO	0.024 ± 0.004	$\mathbf{0.043 \pm 0.003}$

7 Discussion and Conclusion

In this work, we developed hard clustering algorithms that can learn the distance metric from data itself. We achieve this by first showing the connection between Tweedie models and beta divergences. We modified the Bregman hard clustering and DP-Means algorithms to work with beta divergences with the implication that these algorithms are the equivalent of EM for TMMs and DP-TMMs under asymptotically small variance assumptions. We then proposed four new clustering algorithms that can identify the underlying TMM using the population moment conditions. We showed the value of the TMM over simple Gaussian mixture models in simulated data and seven UCI data sets. On a gene expression dataset without labels, we showed that TMM-inspired clustering methods recover latent structure in the presence of unknown and heterogeneous data distributions. In future work, we plan to explore more complicated models with Tweedie building blocks.

References

[1] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

[2] Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

[3] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[4] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[5] Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[6] Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

[7] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. Csiszars divergences for non-negative matrix factorization: Family of new algorithms. In *Independent Component Analysis and Blind Signal Separation*, pages 32–39. Springer, 2006.

[8] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-ichi Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[9] Peter K Dunn and Gordon K Smyth. Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15(4):267–280, 2005.

[10] Peter K Dunn and Gordon K Smyth. Evaluation of tweedie exponential dispersion model densities by fourier inversion. *Statistics and Computing*, 18(1):73–86, 2008.

[11] Christian Gourieroux, Alain Monfort, and Alain Trognon. Pseudo maximum likelihood methods: Theory. *Econometrica: Journal of the Econometric Society*, pages 681–700, 1984.

[12] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

[13] Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

[14] Romain Hennequin, Bertrand David, and Roland Badeau. Beta-divergence as a subclass of bregman divergence. *IEEE Signal Processing Letters*, 18(2):83–86, 2011.

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

[16] Bent Jorgensen. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 127–162, 1987.

[17] Brian Kulis and Michael I Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 513–520, 2012.

[18] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter ACt Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[19] M. Lichman. UCI machine learning repository, 2013.

[20] Minami Mihoko and Shinto Eguchi. Robust blind source separation by beta divergence. *Neural computation*, 14(8):1859–1886, 2002.

[21] Nurul Haque Mollah, Nayeema Sultana, Mihoko Minami, Shinto Eguchi, et al. Robust extraction of local structures by the minimum β-divergence method. *Neural Networks*, 23(2):226–238, 2010.

[22] S Subhash. Genescf: Gene set clustering based on functional annotation. `https://github.com/santhilalsubhash/geneSCF.git`, 2014.

[23] Vincent YF Tan and Cédric Févotte. Automatic relevance determination in nonnegative matrix factorization with the/spl beta/-divergence. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(7):1592–1605, 2013.

[24] Y Kenan Yilmaz and A Taylan Cemgil. Alpha/beta divergences and tweedie models. *arXiv preprint arXiv:1209.4280*, 2012.