

# Bhattacharyya clustering with applications to mixture simplifications

Frank Nielsen, Sylvain Boltz, and Olivier Schwander  
*LIX, Ecole Polytechnique - Sony CSL*  
*nielsen,boltz,schwander@lix.polytechnique.fr*

## Abstract

*Bhattacharyya distance (BD) is a widely used distance in statistics to compare probability density functions (PDFs). It has shown strong statistical properties (in terms of Bayes error) and it relates to Fisher information. It has also practical advantages, since it strongly relates on measuring the overlap of the supports of the PDFs. Unfortunately, even with common parametric models on PDFs, few closed-form formulas are known. Moreover, the BD centroid estimation was limited to univariate gaussian PDFs in the literature and no convergence guarantees were provided. In this paper, we propose a closed-form formula for BD on a general class of parametric distributions named exponential families. We show that the BD is a Burbea-Rao divergence for the log normalizer of the exponential family. We propose an efficient iterative scheme to compute a BD centroid on exponential families. Finally, these results allow us to define a Bhattacharyya hierarchical clustering algorithms (BHC). It can be viewed as a generalization of k-means on BD. Results on image segmentation shows the stability of the method.*

## 1 Introduction

BD is a popular distance to compare distributions. It is used for a wide class of applications including speech recognition [5] or various image processing and computer vision tasks such as video tracking [2]. This popularity comes from its simplicity and from its discriminative power. Indeed minimizing BD on some space of distributions strongly enforces overlap between the supports of distributions. On a more theoretical point of view, BD relates to the sup and the inf. on Bayes error [7], while there are no such results for another popular distance the symmetric Kullback-Leibler (SKL). BD and SKL also relates to the Fisher information.

There are two popular techniques in statistics to estimate PDFs from data points, parametric estimation,

praised for its simplicity and nonparametric estimation, praised for its accuracy. A wide class of parametric models on PDFs can be written in a unified manner as exponential families [9]. The class of exponential families contains many of the standard parametric models including Poisson, Gaussian, Multinomial distributions.

However, only a few closed-form formulas for BD between those PDFs are known in the literature. For instance, the BD between multivariate normal distributions is given here [3]. Moreover, the BD centroid defined as the minimizer of a sum of BD is still not well studied, yet very useful point in clustering. Indeed, many clustering algorithms, such as k-means Lloyd Algorithms or Mean-Shift Algorithm [3, 2]. assume that a mean can be easily computed. The BD centroid has been only studied for univariate Gaussian distributions [11], it is shown that it can be estimated as an iterative algorithm, although no convergence guarantees are given. In this paper we study the BD centroid estimation on exponential families for clustering applications. First contribution, a closed-form formula for BD on exponential families is given in Section 2. Second an iterative scheme is computed for any members of exponential families in Section 3, including the multivariate normals. It relies on the ConCave-Convex Procedure (CCCP) [13] and CCCP convergence properties were recently studied [12]. Third, we extend previous work on BD centroids [11] to multivariate Gaussian in order to compare it with our new scheme. Last contribution, we plug those BD centroids into a hierarchical clustering algorithms to define a new Bhattacharyya hierarchical clustering technique (BHC) in Section 4. This BHC is finally studied on image segmentation problems in Section 5.

## 2 Bhattacharyya distance for exponential families

Many usual statistical parametric distributions  $p(x; \lambda)$  (e.g., Gaussian, Poisson, Bernoulli/multinomial, Gamma/Beta, etc.) share

common properties arising from their common canonical decomposition of probability distribution:

$$p(x; \lambda) = \exp(\langle t(x), \theta \rangle - F(\theta) + k(x)). \quad (1)$$

An exponential family is characterized by its *log-normalizer*  $F(\theta)$ , and a distribution in that family by its *natural parameter*  $\theta$  belonging to the *natural space*  $\Theta$ .  $t(x)$  is the sufficient statistic and  $k(x)$  is the carrier measure.

For arbitrary distributions  $p$  and  $q$ , the Bhattacharyya coefficient  $B_c(p, q) = \int \sqrt{p(x)q(x)}dx$  measures the amount of overlap of these distributions. The BD is derived from its coefficient as  $B(p, q) = -\ln B_c(p, q)$ .

Although the distance is symmetric, it is not a metric. Nevertheless, it can be metrized using the Hellinger/Matusita metric  $H(p, q) = \sqrt{1 - B(p, q)}$ . In addition, note that the Voronoi diagram of Helling/Matusita coincide with the one of the BD.

For distributions belonging to the same exponential family, it turns out that the BD is always in closed-form. Namely, it is a Jensen-Shannon divergence of the cumulant function on natural parameters. Indeed, we have

$$\begin{aligned} B_c(p, q) &= \int \sqrt{p(x)q(x)}dx \\ &= \exp\left(\left\langle t(x), \frac{\theta_p + \theta_q}{2} \right\rangle - \frac{F(\theta_p) + F(\theta_q)}{2} + k(x)\right) \\ &= \exp\left(F\left(\frac{\theta_p + \theta_q}{2}\right) - \frac{F(\theta_p) + F(\theta_q)}{2}\right) \geq 0 \end{aligned}$$

It follows that the BD for members of the same exponential family is  $\frac{F(\theta_p) + F(\theta_q)}{2} - F\left(\frac{\theta_p + \theta_q}{2}\right)$ , namely a Burbea-Rao (extending Jensen-Shannon divergence) of the cumulant function on the natural parameters.

The closed-form formula for a particular member of an exponential family is then given by choosing the  $F$  [9]. Some examples are provided in Table 1.

### 3 Method 1: Bhattacharyya centroids for exponential families

Traditionally in the Euclidean space, the centroid is defined as the minimizer of the average sum of squared Euclidean distances, and the median as the minimizer of the Euclidean distances<sup>1</sup>. Because Hellinger distance  $H(p, q) = \sqrt{1 - B(p, q)}$  is a metric built on the

<sup>1</sup>the squared Euclidean distance is not a metric since it violates the triangle inequality

square root of the BD (and that the BD is not a metric), it is natural to call the Bhattacharyya centroid the minimizer of the average BD, called the loss function  $L(c)$ .

The BD centroid is defined as

$$c = \arg \min L(c) = \arg \min \sum_{i=1}^n B(p_i, c) \quad (2)$$

This is in accordance with the Bregman centroid [8] that is an average Bregman divergence minimizer<sup>2</sup>. Since BD is symmetric, right and left sided centroids are equal.

The CCCP [13] is a general purpose loss function minimizer. Provided that the Hessian of the loss function is bounded, we can always decompose an arbitrary (non-convex) function as the sum of a convex function with a concave function:  $L(c) = L_{\text{convex}}(c) + L_{\text{concave}}(c)$ . For the Bhattacharyya centroid, this decomposition is explicit as the term  $L_{\text{convex}}(c) = \frac{F(c)}{2}$  is the convex function and the term  $L_{\text{concave}}(c) = -\sum_{i=1}^n F\left(\frac{p_i + c}{2}\right)$  the concave function (since the sum of concave functions is a concave function). Thus we initialize the parameter  $\Theta_0$  as the SKL centroid (known in closed-form, [8]) and iteratively update

$$\nabla L_{\text{convex}}(\Theta^{(k+1)}) = -\nabla L_{\text{concave}}(\Theta^{(k)})$$

This optimization scheme monotonically decreases the loss function until it reaches a local minimum or saddle point.

$$\Theta^{(k+1)} = \nabla L_{\text{convex}}^{-1}\left(-\nabla L_{\text{concave}}(\Theta^{(k)})\right)$$

We have,

$$\Theta^{(k+1)} = \nabla F^{-1}\left(\frac{1}{n} \sum_i \nabla F\left(\frac{p_i + \Theta^{(k)}}{2}\right)\right) \quad (3)$$

This is a generalized  $f$ -mean [8]. This formula is the main result of the paper, it is an iterative scheme to compute a Bhattacharyya centroid. It has two major improvements on the scheme proposed on univariate Gaussians [11] First, it can be computed on any member of exponential families. Second convergence properties of CCCP have been studied [13, 12]. Note that the CCCP algorithm can also be used to optimize the SKL divergence (that is also known as the Jensen-Shannon divergence). The SKL centroid is a symmetrized Bregman centroid for exponential families [8].

<sup>2</sup>the squared Euclidean distance is a Bregman divergence

Exponential family	$\exp\left(F\left(\frac{\theta_p+\theta_q}{2}\right) - \frac{F(\theta_p)+F(\theta_q)}{2}\right)$
Multinomial	$-\ln \sum_{i=1}^d \sqrt{p_i q_i}$
Poisson	$\frac{1}{2}(\sqrt{\mu_p} - \sqrt{\mu_q})^2$
Gaussian	$\frac{1}{4} \frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2} + \frac{1}{2} \ln \frac{\sigma_p^2 + \sigma_q^2}{2\sigma_p \sigma_q}$
Multivariate Gaussian	$\frac{1}{8}(\mu_p - \mu_q)^t \left(\frac{\Sigma_p + \Sigma_q}{2}\right)^{-1} (\mu_p - \mu_q) + \frac{1}{2} \ln \frac{\det \frac{\Sigma_p + \Sigma_q}{2}}{\det \Sigma_p \det \Sigma_q}$

**Table 1.** BD given in closed-form for some exemplar members of exponential families

## 4 Method 2: Bhattacharyya centroids on multivariate normals

A BD centroid for univariate Gaussians has previously been proposed in [11]. In order to compare it on multivariate data with the scheme proposed in the previous section, we extend this framework to multivariate Gaussians.

The loss function (2) now reads

$$L(c) = \sum_{i=1}^n \frac{1}{8} (\mu_c - \mu_i)^\top \left(\frac{\Sigma_c + \Sigma_i}{2}\right)^{-1} (\mu_c - \mu_i) + \frac{1}{2} \log \left(\frac{\det \left(\frac{\Sigma_c + \Sigma_i}{2}\right)}{\sqrt{\det \Sigma_c \det \Sigma_i}}\right). \quad (4)$$

In order to minimize it, let us differentiate with respect to  $\mu_c$ , let us note  $U_i = (\Sigma_c + \Sigma_i)^{-1}$ . We used [10] p.10 (73)

$$\frac{\partial L}{\partial \mu_c} = \sum_{i=1}^n [U_i + U_i^\top] [\mu_c - \mu_i] \quad (5)$$

Then one can estimate iteratively  $\mu_c$ , since  $U_i$  depends on  $\Sigma_c$  which is unknown.

$$\mu_c(t+1) = \left[ \sum_{i=1}^n [U_i + U_i^\top] \right]^{-1} \left[ \sum_{i=1}^n [U_i + U_i^\top] \mu_i \right] \quad (6)$$

Now let us estimate  $\Sigma_c$ . We used [10] p.9 (55) for the first term, [10] p.8 (51) for the two others.

$$\frac{\partial L}{\partial \Sigma_c} = \sum_{i=1}^n -U_i^\top (\mu_c - \mu_i) (\mu_c - \mu_i)^\top U_i^\top + 2 \sum_{i=1}^n U_i^\top - \sum_{i=1}^n \Sigma_c^{-\top}. \quad (7)$$

Taken into account the fact that  $\Sigma_c$  is symmetric, differential calculus on symmetric matrices can be simply estimate:

$$\frac{dL}{d\Sigma_c} = \frac{\partial L}{\partial \Sigma_c} + \left[ \frac{\partial L}{\partial \Sigma_c} \right]^\top - \text{diag} \left( \frac{\partial L}{\partial \Sigma_c} \right). \quad (8)$$

Thus, if one notes

$$A = \sum_{i=1}^n 2U_i^\top - U_i^\top (\mu_c - \mu_i) (\mu_c - \mu_i)^\top U_i^\top \quad (9)$$

and recalling that  $\Sigma_c$  is symmetric, one has to solve

$$n(2\Sigma_c^{-1} - \text{diag}(\Sigma_c^{-1})) = A + A^\top - \text{diag}(A). \quad (10)$$

Thus if one notes

$$B = A + A^\top - \text{diag}(A) \quad (11)$$

Then one can estimate  $\Sigma_c$  iteratively.

$$\Sigma_c^{(k+1)} = 2n \left[ (B^{(k)} + \text{diag}(B^{(k)})) \right]^{-1} \quad (12)$$

Finally, in this paper we propose two methods to compute a centroid that minimizes a BD loss function. Method 1 initializes with SKL centroid [8], and updates it with the general iterative scheme for Exponential Families presented in Eq.(3). Method 2 initializes with SKL centroid and updates it alternatively with Eq.(12) and Eq.(6). With these two methods for computing centroids, one can do Bhattacharyya Hierarchical Clustering (BHC), building on Bregman Hierarchical Clustering described in [4] by plugging-in our techniques to compute BD centroids and closed-form BD.

## 5 Experiments

The first results presented on Fig.1 are visual, the goal is here to demonstrate the stability of our clustering. The image features used in this experiment are 5 dimensional (joint color and position). Note that method 1, Eq.(3), to compute BDs and BD centroids was used here. Comparisons with other Gaussian mixtures simplification techniques are out of the scope of this paper. However, results on the same images with Bregman hierarchical clustering (using SKL distance) can be found on the jMEF home page [9]. In particular, BHC seems to perform better on the last image. The second experiment is numerical. Since we present two different



**Figure 1.** Segmentation results, first row : 4 Images, second row 4 images segmented with a mixture of 48 5D Gaussians using hierarchichal mixture model, third row, Gaussian mixtures reduced to 16 5D Gaussians using BHC

schemes to compute the BD centroid, one wants to compare them in terms of stability and accuracy. Whenever the ratio of BD loss function between those centroids is greater than 1%, we consider that one of the two algorithms has failed (the one that gives the highest BD). Among 760 centroids computed to generate Fig.1, 100% were correct with method 1, and 87% were correct with method 2. The average number of iterations to reach this 1% accuracy is 4.1 for method 1, and 5.2 for method 2. In terms of generalization (method 1 is valid for all exponential families), stability (100% of relative convergence rate) and speed (1 iteration per centroid difference) , method 1, Eq.(3), seems to provide the most efficient scheme to compute BD centroids.

## 6 Conclusion

In this paper, we have shown how BD can be computed in closed-form for exponential families. From this formula, we were able to derive an efficient iterative scheme to compute a BD centroid that seems to perform better than an existing iterative scheme. This scheme has proven to be simple yet very efficient to compute the generalized k-means algorithms on BDs. An implementation of this clustering algorithm will be provided in the jMEF library. Future works will focus on BD and BD centroids over mixture of exponential families. It seems that BD on mixtures can be accurately estimated from the BD between single components of the mixture [5]. Finally, connections with Riemannian geometry and optimization on the cone of positive definite matrices [9] will be studied.

## Acknowledgments

We gratefully acknowledge financial support from DIGITEO GAS 2008-16D and ANR GAIA 07-BLAN-0328-01.

## References

- [1] A. Acero and M. D. Plumpe. Method for training of subspace coded gaussian models, 2009. US PATENT 7571097.
- [2] D. Comaniciu, V. Ramesh, and P. Meer. Real-time tracking of non-rigid objects using mean shift. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 142–151, Hilton Head Island, SC, 2000.
- [3] K. Fukunaga. *Introduction to statistical pattern recognition (2nd ed.)*. Academic Press Professional, Inc., 1990.
- [4] Vincent Garcia, Frank Nielsen, and Richard Nock. Levels of details for gaussian mixture models. In *Asian Conference on Computer Vision (ACCV)*, Xi an, China, September 2009.
- [5] J. Hershey and P. Olsen. Variational bhattacharyya divergence for hidden markov models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.
- [6] T. Jebara and R. Kondor. Bhattacharyya and expected likelihood kernels. In *Learning theory and Kernel machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003: proceedings*, page 57. Springer Verlag, 2003.
- [7] T. Kailath. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Transactions on Communication Technology*, 15(1):52–60, 1967.
- [8] F. Nielsen and R. Nock. Sided and symmetrized bregman centroids. *IEEE Transactions on Information Theory*, 55(6):2048–2059, 2009.
- [9] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards, 2009.
- [10] K. B. Petersen and M. S. Pedersen. *The Matrix Cookbook*. Technical University of Denmark, oct 2008.
- [11] L. Rigazio, B. Tsakam, and J.C. Junqua. Optimal Bhattacharyya centroid algorithm for Gaussian clustering with applications in automatic speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1599–1602, 2000.
- [12] Bharath Sriperumbudur and Gert Lanckriet. On the convergence of the concave-convex procedure. In *Neural Information Processing Systems*, 2009.
- [13] AL Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15(4):915–936, 2003.