CLOSED-FORM CAUCHY-SCHWARZ PDF DISTANCE FOR MIXTURE OF GAUSSIANS

Kittipat Kampa and Jose C. Principe

Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611

ABSTRACT

This paper presents an efficient approach to calculate the difference between two probability density functions (pdfs), each of which is a mixture of Gaussians (MoG). Unlike Kullback-Leibler divergence (D_{KL}), the authors propose that the Cauchy-Schwarz (CS) pdf divergence measure (D_{CS}) can give an analytic closed-form expression for MoG. This property of the D_{CS} makes fast and efficient calculations possible, which is desired tremendously in real-world applications where the dimensionality of the data/features is very high. We show that D_{CS} follows similar trends as D_{KL}, but can be computed much faster, especially when the dimensionality is high. Moreover, the proposed method is shown to significantly outperform D_{KL} in classifying real-world 2D and 3D objects based on distances alone.

Index Terms— Cauchy-Schwarz, pdf distance, Kullback-Leibler divergence, closed-form solution, mixture of Gaussians, 2D and 3D object recognition

1. INTRODUCTION

The Gaussian mixture model has been a very useful probability model for a variety of applications due to the fact that the number of parameters used in a mixture of Gaussians (MoG) is very small and due to its flexibility to model distributions whose parametric forms are unknown. In many applications, one would like to compare two pdfs, each of which is a MoG, by measuring the difference between the two pdfs using various types of available divergences or distances measures. However, not all divergences are equally useful for the MoG model because most well known divergences, including the Kullback-Leibler divergence (D_{KL}), do not yield an analytic closed-form expression for MoG.

To work around this problem, a few approaches are used to estimate the D_{KL} in practice, such as numerical integration (NI) and stochastic integration (SI) [1, 2]. In NI, the whole feature space is uniformly gridded, then each gridded cell is used in the calculation of the D_{KL}. Thus, the accuracy will highly depend on the resolution of the grid. The smaller the grid size, the better accuracy obtained, but this comes with the cost of a larger memory size being used to store those cells. This is the trade-off between memory and accuracy when using NI. The size of memory used in NI grows exponentially with dimensionality of the data/feature vector (the curse of dimensionality). Another big drawback of NI is that it sometimes misses narrow peaks of MoG components [3]. Stochastic integration techniques have been proposed to mitigate this problem by sampling from the MoG of interest, which lessens the chance of missing narrow peaks. However, this approach cannot avoid the dimensionality curse because when the dimensionality increases, the number of samples has to increase in order to keep up with the additional details of the MoG. Also, there is no theoretical criterion to relate the number of samples with the dimensionality. Consequently, a closed-form expression for divergence of a MoG model is desired.

The organization of this paper is as follows. In section 2, the authors investigate why D_{KL} does not yield a closed-form expression for MoG. Next, in section 3, the author show that the Cauchy-Schwarz (CS) pdf divergence measure yields an analytic closed-form expression for MoG, eliminating the problems of grid size and scaling. Numerical examples are shown in section 4. Finally we test the proposed expression in real-world 2D and 3D object classifications in section 5. We found that D_{CS} has similar behavior to that of D_{KL}, but D_{CS} has advantages in computational time which could lead to real-time application or machine learning on mobile devices.

2. CLOSED-FORM EXPRESSION OF D_{KL} FOR MOG?

In this section, we demonstrate that the analytic closed-form expression of D_{KL} for MoG is not possible. Let $p(x)$ and $q(x)$ denote two distributions, each of which is a mixture of Gaussians with different parameters and number of clusters: $q(x) = \sum_{m=1}^{M} \pi_m \mathcal{N}(x|\mu_m, \Lambda_m^{-1})$ and $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Lambda_k^{-1})$ where M and K denote the number of Gaussian components in $q(x)$ and $p(x)$ respectively. Let π_i , μ_i and Λ_i denote the mixture coefficient, the mean and the precision matrix of the i^{th} component of an MoG respectively. The multivariate Gaussian distribution is given by $\mathcal{N}(x|\mu_i, \Lambda_i^{-1}) = \frac{|\Lambda_i|^{1/2}}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^\top \Lambda_i (x-\mu_i)\right)$ where $x \in R^D$. The integral and the summations in D_{KL} are separated by the log which disallows the marginalization

This work is partially funded by ONR N00014-10-1-0375.

of x out by the integral at the front, preventing an analytic solution.

The authors also investigate other divergences with the integral inside the log in the hope that the integral would be distributed inside the summation and, thus, x will be marginalized out. The α-divergence [4] is one well-known measure, however, it failed to give a closed form expression despite having the integral inside the log. That is because there is not clear insight into how to calculate an x-independent closed-form expression from the inverse of a MoG generated by a minus-sign power $(1-\alpha)$ in the α-divergence.

From the failures learned from both divergences and Gaussian identities, we can reduce our search domain for a distance/divergence significantly. The authors found that a closed-form expression for MoG can be derived from some divergences such as the Cauchy-Schwarz pdf divergence measure (D_{CS}) [5], Jensen-Renyi divergence (D_{JR}) [6] and corresponding concordance [7]. In this paper, we restrict our discussion to D_{CS} as it has behavior similar to D_{KL} [8] and it is simple to implement and to understand due to its relation to the Cauchy-Schwarz inequality.

3. CLOSED-FORM EXPRESSION FOR D_{CS}

Inspired by the renowed Cauchy-Schwarz inequality, the Cauchy-Schwarz PDF divergence measure [5] is given by:

$$D_{CS}(q,p) = -\log\left(\frac{\int q(x)p(x)dx}{\sqrt{\int q(x)^2dx \int p(x)^2dx}}\right). \quad (1)$$

This is a symmetric measure for any two pdfs q and p, such that $0 \leq D_{CS} < \infty$, where the minimum is obtained if and only if $q(x) = p(x)$. The measure plays important roles in information theoretic learning (ITL), non-parametric density estimation [9], graph theory, Mercer kernel theorey and spectral theory [5]. Before we show its derivation, it is important to see why D_{CS} is considered in the MoG case. At first glance, it is obvious that the integral can be distributed into the weighted summation of Gaussian components because every thing is now inside the log. Also we envision that the integral of multiplication of MoGs is a MoG in the space of the mean parameters μ. In this section, the closed-form expression for MoG is derived as follows. Equ. 1 is rewritten as $D_{CS}(q,p) = -\log\left(\int q(x)p(x)dx\right) + \frac{1}{2}\log\left(\int q(x)^2dx\right) + \frac{1}{2}\log\left(\int p(x)^2dx\right)$. By distributing the integral into the summation, and using the Gaussian identity, the first term on the r.h.s., $\log\left(\int q(x)p(x)dx\right)$, can be written in a closed-form expression independent of x:

$$\log\left(\int \sum_{m=1}^{M}\sum_{k=1}^{K} \pi_m\pi_k \mathcal{N}(x|\mu_m,\Lambda_m^{-1})\mathcal{N}(x|\mu_k,\Lambda_k^{-1})dx\right)$$
$$= \log\left(\sum_{m=1}^{M}\sum_{k=1}^{K} \pi_m\pi_k \int \mathcal{N}(x|\mu_m,\Lambda_m^{-1})\mathcal{N}(x|\mu_k,\Lambda_k^{-1})dx\right)$$
$$= \log\left(\sum_{m=1}^{M}\sum_{k=1}^{K} \pi_m\pi_k z_{mk}\right).$$

dimensionality (D)	D_{CS}	$D_{KL}^{(NI)}$	$D_{KL}^{(SI)}$
2	0.002	26	4.3
3	0.002	37.2	4.3

Table 1. Average run-time (seconds) in numerical experiment

Applying the same trick to the second and third term on the r.h.s., the closed-form expression is given by:

$$\begin{aligned} &D_{CS}(q,p) = \\ &-\log\left(\sum_{m=1}^{M}\sum_{k=1}^{K} \pi_m\pi_k z_{mk}\right) \\ &+\frac{1}{2}\log\left(\sum_{m=1}^{M} \frac{\pi_m^2|\Lambda_m|^{1/2}}{(2\pi)^{D/2}} + 2\sum_{m=1}^{M}\sum_{m'<m} \pi_m\pi_{m'} z_{mm'}\right) \\ &+\frac{1}{2}\log\left(\sum_{k=1}^{K} \frac{\pi_k^2|\Lambda_k|^{1/2}}{(2\pi)^{D/2}} + 2\sum_{k=1}^{K}\sum_{k'<k} \pi_k\pi_{k'} z_{kk'}\right) \end{aligned} \quad (2)$$

where $z_{mk} = \mathcal{N}(\mu_m|\mu_k,(\Lambda_m^{-1}+\Lambda_k^{-1}))$. The expression has a complexity of order $O(M^2)$ when $M \geq K$, which is much smaller than that of NI and SI whose complexities depend on the dimensionality of the data D and the number of sample is $N \gg M^2$ in general.

4. NUMERICAL EXAMPLES

In this experiment, the behavior of D_{CS} and that of D_{KL} are compared in terms of their values in several circumstances and their run-time when the feature vectors are all 2D. The parameters for $q(x) = q([x_1\, x_2]^\top)$ are given by $M = 3$, $(\pi_1,\mu_1,\Sigma_1) = (0.3,[0\ 0]^\top, I)$, $(\pi_2,\mu_2,\Sigma_2) = (0.3,[3\ 0]^\top, I)$, $(\pi_3,\mu_3,\Sigma_3) = (0.4,[8\ 0]^\top, I)$. For illustrative purposes, the distribution $p(x)$ is picked to be an x_2-shifted version of $q(x)$, that is $p(x) = q([x_1\, x_2+\Delta x_2]^\top)$ where the shifts are $\Delta x_2 \in \{-20,-19,...,19,20\}$. In this experiment, we calculate divergences using 3 approaches: 1) D_{CS}, 2) KL-divergence using NI ($D_{KL}^{(NI)}$) and 3) KL-divergence using SI ($D_{KL}^{(SI)}$). The D_{CS} is calculated by Equ. 2. The $D_{KL}^{(NI)}$ is evaluated on the region of interest (x_1,x_2): $x_1 = [-15,23]$ and $x_2 = [-15,15]$ with equal resolution of 0.01 on both the x_1 and x_2 axis. The $D_{KL}^{(SI)}$ is evaluated by sampling (like in Monte Carlo method) from the distributions $q(x)$ using number of sample $N = 1000$ samples. The results and run-time are shown in Fig. 1 and Table 1 respectively.

The results in Fig. 1 show that when $q(x)$ and $p(x)$ are the same distribution, the divergence $D_{CS} = 0$, and when they are moving away from each other, the value of D_{CS} is

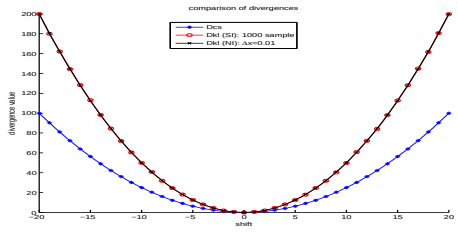


Fig. 1. Comparison of divergences evaluated from D_{CS}, $D_{KL}^{(NI)}$ and $D_{KL}^{(SI)}$ in numerical experiment.

increasing. Furthermore, the curve shows the similarity of the behavior between D_{CS} and D_{KL} in terms of performance. The average run-time is shown in Table 1. Additionally, we perform a similar experiment in 3D vector space, and found that when the dimensionality of the data increases, the run-time of $D_{KL}^{(NI)}$ increases significantly, whereas that of D_{CS} remains almost the same. The details will be discussed more in a subsequent section. In the next section, the divergences are put to the test on real-world 2D and 3D object classification.

5. OBJECT CLASSIFICATION

Object classification has been an active area of research for years. Nowadays, the area of object classification is more interesting and challenging because richer and more informative feature vectors can be obtained by novel sensor technology, which in turn makes it possible to classify objects of more complicated shapes in many real-world applications. Several approaches have been proposed. Pdf-based classification [8, 10, 11] is among the well-known methods, as the pdf of features can be viewed as a very informative descriptor of an object [8]. In this paper, we take advantage of having extracted features as MoGs to exploit pdf-distance-based classification, i.e. the model/class whose divergence to the test sample is smallest will be assigned to the sample. Let $\mathcal{C}$ denote the set of models, $c_i \in \mathcal{C}$ denote the i^{th} model, s_j denote the j^{th} (test) sample, $q(x|s_j)$ and $p(x|c_i)$ denote the pdf of extracted features for s_j and c_i respectively. Then the class $c^\star$ will be assigned to the sample s_j if $c^\star = \arg\min_{c_i \in \mathcal{C}} D(q(x|s_j)||p(x|c_i))$. In this section we present the performance of the proposed algorithms in 2D and 3D object recognition.

5.1. Experiment1: Object classification in 2D

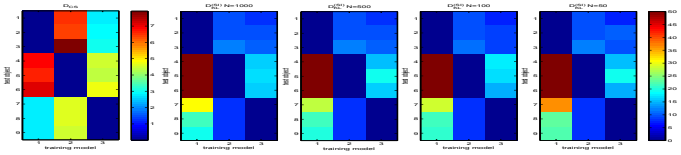
In this experiment we have 3 types/classes of images (front cover of 3 books), A, B and C as shown in Fig. 2 (a). The features are extracted from each image by converting RGB value of each pixel to CIELuv [12] which is a very powerful feature in image recognition because the Euclidean distance

D_{CS}	D_{KL}^{1000}	D_{KL}^{500}	D_{KL}^{100}	D_{KL}^{50}
0.006	3.7	1.9	0.37	0.19

Table 2. Run-time of experiment1

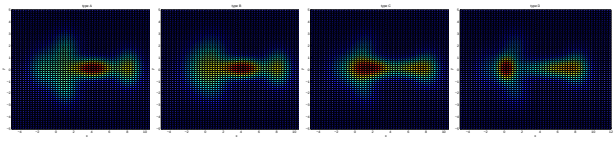
between two sets of color coordinates approximates the human perception of color difference. Therefore, each image is modeled by a 3D MoG with 2 Gaussian components decided using BIC. The probability model of each type is built using 4 sample images of the same book with different scales and orientations. The performance of D_{CS} in Equ. 2 is tested against that of D_{KL} (with varying number of samples N = 1000, 500, 100, 50) on the test datasets containing 3 types of front covers, each of which has 3 sample images, so we will have 9 test images total. The objective is to classify the front cover of 3 books based on the pdf distance mentioned earlier. The divergences evaluated in each case are shown in Fig. 2 (b) and will be discussed at the end of this section.

(a) Sample of type A, B and C from left to right respectively



(b) (left-most) The results from using D_{CS}. Next are D_{KL} using number of sample N= 1000, 500, 100, 50 respectively.

Fig. 2. Data samples and results from Experiment1.



(a) Type A (b) Type B (c) Type C (d) Type D

Fig. 3. 2D footprints of objects in Experiment2.

5.2. Experiment2: Object classification in 3D

In this experiment we have 4 types/classes of synthesized 3D objects A, B, C and D whose 2D footprints are shown in Fig. 3 (a), (b), (c) and (d), each of which is modeled by 3D-MoG with 6, 6, 7 and 7 Gaussian components respectively. The inputs of this experiment are MoG parameters learned from the object point cloud. We test the performance of D_{CS} against

	noise level			run-time
	ns1	ns2	ns3	(sec)
D_{CS}	100	100	82.5	0.037
$D_{KL}^{(N=1000)}$	100	87.5	65	5.82
$D_{KL}^{(N=500)}$	100	83.75	60	2.92
$D_{KL}^{(N=200)}$	98.33	80.83	55	1.15
$D_{KL}^{(N=100)}$	96.87	78.12	56.88	0.58

Table 3. Total percent accuracy matrix of Experiment2

D_{KL} (with varying number of samples N = 1000, 500, 200, 100) on 3 datasets; Dataset_ns1, ns2 and ns3–the number represents the noise level from low to high. Each dataset contains 4 object types and each type has 10 samples, so we will have 40 samples total in each dataset. Our goal is to classify/label each sample based on minimum divergence criteria mentioned earlier. The performances of D_{CS} and $D_{KL}^{(SI)}$ are shown in the total accuracy matrix in Table 3.

5.3. Discussion of object classification

Since, by the nature of this dataset, the complexity of distributions in Experiment3 is low so it can be fitted appropriately using an MoG with 2 components. That yields the 100% classification accuracy for all the approaches as shown in Fig. 2 (b), and more interestingly, D_{CS} tends to be more time-efficient as shown in Table. 2.

The results in Experiment2 indicate excellent performance of both approaches in the noise-free environment (ns1). In intermidiate noise level (ns2), the D_{CS} still performs flawlessly, but $D_{KL}^{(SI)}$ performance decreases as the number of samples decreases. The same behavior of $D_{KL}^{(SI)}$ also appears when the noise level is high (ns3), in which case the performance of D_{CS} also slightly drops. By all the cases, D_{CS} still outperforms D_{KL}.

The results from Experiment1 and 2 show that the closed-form expression of D_{CS} outperforms and computationally outruns that of D_{KL} significantly in both applications. That is because the number of parameters used by an MoG is only $M(\frac{D}{2}+1)(D+1)$, which is much less than the number of samples N needed in order to maintain a good estimator of a distribution. Furthermore, when the number of dimensions D goes very high, the sample size N in D_{KL} will grow exponentially with the dimension, whereas, the closed-form expression in D_{CS} only depends on the number of mixture components. The results also illustrate similar trends of D_{CS} and $D_{KL}^{(SI)}$ which implies that replacing $D_{KL}^{(SI)}$ with D_{CS} is possible in many applications especially when the input is given in terms of MoG, when fast computation is required, and when there are limitations in computational resources and power consumption which usually happens when working in modern mobile or hand-held devices.

6. CONCLUSION AND FUTURE WORK

In this paper we illustrate why D_{KL} and α-divergence do not give a closed-form expression for MoG. From this observation, we come up with some preliminary criteria to search for such divergences, however, we restrict our attention to the Cauchy-Schwarz pdf diverge measure [5]. Using the Gaussian multiplication identity, we come up with a closed-form expression for D_{CS} which does not depend on x. This reduces the complexity to $O(M^2)$ when $M \geq K$, which is much smaller than that of NI and SI whose complexities depend on the number of samples $N \gg M^2$ in general.

We also show that D_{CS} outperforms and outruns $D_{KL}^{(SI)}$ significantly in real-world object classification in both 2D and 3D. Additionally, similar trends between both approaches suggest the possibility to replace D_{KL} with D_{CS} in many real-world applications where the form of input is appropriate. In future work, we will further pursue more general criteria to pinpoint such divergences in the hope that the criteria will lead to the way to construct divergences that give a closed-form expression for MoG.

7. REFERENCES

[1] Jesper Hjvang, Jensen Daniel, P. W. Ellis, Mads G. Christensen, and Sren Holdt Jensen, "Evaluation of distance measures between gaussian mixture models of mfccs," 2007.

[2] Elias Pampalk, "Speeding up music similarity," in *in Proceedings of the MIREX Annual Music Information Retrieval eXchange*, 2005.

[3] David Mackay, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press, 2005.

[4] Alfred O. Hero, Bing Ma, Olivier Michel, and John Gorman, "Alpha-divergence for classification, indexing and retrieval," Tech. Rep., Communication and Signal Processing Laboratory, Technical Report CSPL-328, U. of Mich., 2001.

[5] Robert Jenssen, Deniz Erdogmus, Kenneth, José C. Príncipe, and Torbjørn Eltoft, "Optimizing the cauchy-schwarz pdf distance for information theoretic, non-parametric clustering.," in *EMMCVPR*, 2005, pp. 34–45.

[6] A. Ben Hamza and H. Krim, "Jensen-renyi divergence measure: theoretical and computational perspectives," jun. 2003, pp. 257 – 257.

[7] Surajit Ray, *Distance-based Model-Selection with application to Analysis of Gene Expression Data*, Ph.D. thesis, 2003.

[8] Jose Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, Springer, 2010.

[9] Sudhir Rao, Allan de Medeiros Martins, and José C. Príncipe, "Mean shift: An information theoretic perspective," *Pattern Recogn. Lett.*, vol. 30, no. 3, pp. 222–230, 2009.

[10] Rui Liao, Christoph Guetter, Chenyang Xu, Yiyong Sun, Ali Khamene, and Frank Sauer, "Learning-based 2d/3d rigid registration using jensen-shannon divergence for image-guided surgery," in *Medical Imaging and Augmented Reality*, Guang-Zhong Yang, Tianzi Jiang, Dinggang Shen, Lixu Gu, and Jie Yang, Eds., vol. 4091 of *Lecture Notes in Computer Science*, pp. 228–235. Springer Berlin / Heidelberg, 2006.

[11] A. Mastin, J. Kepner, and J. Fisher, "Automatic registration of lidar and optical images of urban scenes," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 0, pp. 2639–2646, 2009.

[12] Günther Wyszecki and W. S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae (Wiley Series in Pure and Applied Optics)*, Wiley-Interscience, 2 edition, August.