Closed-Form Information-Theoretic Divergences for Statistical Mixtures

Frank Nielsen
Sony Computer Science Laboratories Inc.
http://www.informationgeometry.org

Abstract

Statistical mixtures such as Rayleigh, Wishart or Gaussian mixture models are commonly used in pattern recognition and signal processing tasks. Since the Kullback-Leibler divergence between any two such mixture models does not admit an analytical expression, the relative entropy can only be approximated numerically using time-consuming Monte-Carlo stochastic sampling. This drawback has motivated the quest for alternative information-theoretic divergences such as the recent Jensen-Rényi, Cauchy-Schwarz, or total square loss divergences that bypass the numerical approximations by providing exact analytic expressions. In this paper, we state sufficient conditions on the mixture distribution family so that these novel non-KL statistical divergences between any two such mixtures can be expressed in generic closed-form formulas.

1. Introduction and motivation

An exponential family [4] is a set of parametric probability distributions $\{p_F(x;\theta) \mid \theta \in \Theta\}$ whose probability density (or mass) can be decomposed canonically as $p_F(x;\theta) = e^{\langle t(x),\theta\rangle - F(\theta)+k(x)}$, where $t(x)$ denotes the sufficient statistics, θ the natural parameter, $F(\theta)$ the log-normalizer, and $k(x)$ the auxiliary carrier measure. $\langle x, y\rangle = x^T y$ denotes the inner product of vectors. Let $\Theta = \{\theta \mid \int p_F(x;\theta)\mathrm{d}x < \infty\}$ denote the natural parameter space. It can be proved [4] that the log-normalizer F is a strictly convex and differentiable function on an open convex set Θ. Canonical decompositions for familiar distributions are reported in the next Section.

21st International Conference on Pattern Recognition (ICPR), pp.1723-1726, 11-15 Nov. 2012.

Now, consider $m(x) = \sum_{i=1}^{k} w_i p_F(x;\theta_i)$ and $m(x) = \sum_{i=1}^{k'} w'_i p_F(x;\theta'_i)$ two finite mixtures of the *same* exponential families with k and k' components, respectively (with positive weights $w_i > 0$ and $w'_i > 0$ summing up to one). Mixtures of exponential families are often met in imaging, pattern recognition and computer vision applications: See [10] for Gaussian mixture models (GMMs), [14] for Rayleigh mixture models (RMMs), [1] for Laplacian mixture models (LMMs), [2] for Bernoulli mixture models (BMMs), etc.

A traditional information-theoretic measure of the statistical dissimilarity between two probability models m and m' is the Kullback-Leibler divergence (KL) $\mathrm{KL}(m : m') = \int_{x\in\mathbb{X}} m(x)\log\frac{m(x)}{m'(x)}\mathrm{d}x$. The KL divergence is also called the relative entropy as it can be expressed as the difference between the cross-entropy $H^{\times}(m : m') = \int_{x\in\mathbb{X}} -m(x)\log m'(x)\mathrm{d}x$ minus the entropy $H(m) = H^{\times}(m : m) = \int_{x\in\mathbb{X}} -m(x)\log m(x)\mathrm{d}x$.

When $k = k' = 1$, the mixtures are degenerated to a single component (with $w_1 = w'_1 = 1$), and the KL divergence between two members of the same exponential family admits a closed-form expression [3, 13]:

$$\mathrm{KL}(m : m') = \int_{x\in\mathbb{X}} p_F(x;\theta_1)\log\frac{p_F(x;\theta_1)}{p_F(x;\theta'_1)}\mathrm{d}x, \quad (1)$$

$$= F(\theta'_1) - F(\theta_1) - \langle\theta'_1 - \theta_1, \nabla F(\theta_1)\rangle, \quad (2)$$

$$= B_F(\theta'_1 : \theta_1), \quad (3)$$

where $B_F(\theta'_1 : \theta_1)$ is a Bregman divergence computed on the *swapped* natural parameters [3, 13]. However, whenever $k + k' > 2$, the KL divergence does *not* admit anymore an analytical expression due to the *log-sum terms* [12] in the mathematical expression of the KL divergence between mixtures. One way to circumvent this problem is to *estimate* the KL divergence using stochastic Monte-Carlo sampling [7]:

$$\tilde{\mathrm{KL}}(m:m') \approx \frac{1}{n} \sum_{\substack{x_i \sim m \\ 1 \leq i \leq n}} \left(\log \frac{m(x_i)}{m'(x_i)} + \frac{m'(x_i)}{m(x_i)} - 1 \right) \quad (4)$$

The Monte-Carlo simulation requires in practice a *large sample* (say, million to billion size) to get a close estimate [5], but is much better than a naive numerical integration of the KL divergence that proceeds by discretizing the support $\mathbb{X}$, and is therefore limited to small dimensions of $\mathbb{X}$. Note that we purposely added the terms $\sum_{i \in \{1,\ldots,n\}} \frac{m'(x_i)}{m(x_i)} - 1$ in Eq. 4 that cancels out as the sampling size increases in order to always ensure that $\tilde{\mathrm{KL}} \geq 0$ (discrete Itakura-Saito divergence for the $x_i \sim m$).

Another approach to circumvent the lack of closed-form solution for the KL divergence of mixtures is to seek for novel notions of statistical distances that allow closed-form expression between mixtures, and compare their experimental performance in real-world applications with respect to the gold standard KL divergence. The *Jensen-Rényi divergence* [16] based on *Rényi quadratic entropy* is such a successful example when considering Gaussian mixture models for point set pattern matching in medical applications. Wang et al. [16] proved that the Jensen-Rényi divergence performs experimentally better than the Jensen-Shannon divergence for point set registration. Liu et al. [11] showed that the *total square loss (tSL)* (an example of a total Bregman divergence [11]) admits an analytic expression for GMMs, and use the robust t-center defined as the minimizer of the average tSL to cluster shapes modeled by GMMs for efficient shape retrieval tasks.

Let us consider the Cauchy-Schwarz (CS) divergence that has been recently introduced in [9, 10]:

$$\mathrm{CS}(P:Q) = -\log \frac{\int p(x)q(x)\mathrm{d}x}{\sqrt{\int p(x)^2\mathrm{d}x \int q(x)^2\mathrm{d}x}}, \quad (5)$$

and supported successfully by several information retrieval applications [10]. The underlying idea of the CS divergence is to rely on the Cauchy-Schwarz inequality of the density functions: $0 < \left(\int p(x)q(x)\mathrm{d}x\right)^2 \leq \int p(x)^2\mathrm{d}x \int q(x)^2\mathrm{d}x$.

It follows that $0 < \frac{(\int p(x)q(x)\mathrm{d}x)^2}{\int p(x)^2\mathrm{d}x \int q(x)^2} \leq 1$ with equality if and only if $p(x) = q(x),\ \forall x \in \mathbb{X}$. The Cauchy-Schwarz divergence bears some similarities with the well-known Bhattacharyya divergence:

$$B(p,q) = -\log \int_{x \in \mathbb{X}} \sqrt{p(x)q(x)}\mathrm{d}x \quad (6)$$

and the squared Hellinger distance:

$$H^2(p,q) = \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}x = 2(1 - e^{-B(p,q)}). \quad (7)$$

Both the Bhattacharyya and Helling distances does not admit closed-form expressions when dealing with mixture models. The key concept of the CS divergence is to consider the raw product distribution $m(x)m'(x)$ (instead of their square root) that allows one to *slide* the integral operand inside the product of components as explained in the next section. In [10], a closed-form formula is *manually* derived for Gaussian mixture models. Since Gaussian mixtures are a special case of exponential family mixtures [3, 13] (EFMs), we show how to derive the calculation and prove a sufficient condition on the exponential families to get a *generic closed-form formula*.

2. The CS divergence of EFMs

Consider the mixture product term $\int m(x)m'(x)\mathrm{d}x$, and let us slide the integral operand inside the $k \times k'$ product terms. We get $\int m(x)m'(x)\mathrm{d}x = \sum_{i=1}^{k} \sum_{j=1}^{k'} w_i w'_j \int p_F(x;\theta_i)p_F(x;\theta'_j)\mathrm{d}x$. Let us now give an analytic expression for the expression $\int p_F(x;\theta_i)p_F(x;\theta'_j)\mathrm{d}x$ whenever the natural parameter space Θ is a *cone* (closed under linear combinations with positive coefficients). Wlog., we may assume the auxiliary carrier measure is zero ($k(x) = 0$). Otherwise, we consider the following change of variable: $\mathrm{d}y = e^{k(x)}\mathrm{d}x$ with $y = \int e^{k(x)}\mathrm{d}x = y(x)$ the anti-derivative. For example, consider the Rayleigh distributions: $p(x;\sigma) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$ defined over the support $\mathbb{X} = \mathbb{R}^+$ for $\sigma > 0$. Rewriting the density as $e^{-\frac{x^2}{2\sigma^2} + \log x - \log \sigma^2}$, we get the following canonical decomposition: $t(x) = -x^2/2$, $\theta = 1/\sigma^2$, $k(x) = \log x$ and $F(\theta) = \log \sigma^2 = -\log \theta$. Let $\mathrm{d}y = e^{k(x)}\mathrm{d}x = x\mathrm{d}x$. It follows that $y = \frac{1}{2}x^2$. We can rewrite the Rayleigh distribution as $p(y;\theta) = e^{-y\theta - F(\theta)}$ for y belonging to the support $\mathbb{Y} = \mathbb{R}^+$. The integral of the product of two such components yields $\int p_F(x;\theta_i)p_F(x;\theta'_j)\mathrm{d}x = \int e^{\langle t(x),\theta_i\rangle - F(\theta_i)} e^{\langle t(x),\theta'_j\rangle - F(\theta'_j)}\mathrm{d}x$. This is equal to $\int e^{\langle t(x),\theta_i+\theta'_j\rangle - F(\theta_i) - F(\theta'_j) - F(\theta_i+\theta'_j) + F(\theta_i+\theta'_j)}\mathrm{d}x = e^{F(\theta_i+\theta'_j) - (F(\theta_i)+F(\theta'_j))} \underbrace{\int e^{\langle t(x),\theta_i+\theta'_j\rangle - F(\theta_i+\theta'_j)}\mathrm{d}x}_{=1}$, since $\theta_i + \theta'_j \in \Theta$. If the natural parameter space Θ is not a cone then the integral $\int e^{\langle t(x),\theta_i+\theta'_j\rangle - F(\theta_i+\theta'_j)}\mathrm{d}x$ *does not converge* by definition [4] whenever

$\theta_i + \theta'_j \notin \Theta$.

Therefore let us assume in the remainder that Θ is a convex cone. The integral of the product of mixtures can be expressed in closed-form: $\int m(x)m'(x)\mathrm{d}x = \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j e^{F(\theta_i+\theta'_j)-(F(\theta_i)+F(\theta'_j))} = \sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j e^{\Delta_F(\theta_i,\theta'_j)}$, with $\Delta_F(\theta_i,\theta'_j) = F(\theta_i+\theta'_j)-(F(\theta_i)+F(\theta'_j))$. Plugging this formula into Eq. 5, we get a closed-form expression of the CS divergence. Note that the term Δ_F is symmetric ($\Delta_F(\theta_i,\theta'_j) = \Delta_F(\theta'_j,\theta_i)$) but it is *not* a divergence ($\Delta_F(\theta,\theta) \neq 0$).

This generic formula generalizes the formula for Gaussians formerly reported in [10] (Eq. 3) to arbitrary exponential families that satisfies the fact that the natural parameter space Θ is a cone. In particular, the closed-form expression applies to mixtures of Bernoulli [2], zero-centered Laplacian [1], Wishart [8, 6] and Gaussian distributions [10] among others with the following respective decompositions:

- **Bernoulli.** $p(x;\lambda) = \lambda^x(1-\lambda)^{1-x}$ (with $\lambda \in (0,1)$), $\theta = \log\frac{\lambda}{1-\lambda}$, $\Theta = \mathbb{R}$, $F(\theta) = \log(1+e^\theta)$.

$$\Delta_{\mathrm{Bernoulli}}(\lambda_i,\lambda_j) = \log\frac{1+\frac{\lambda_i+\lambda_j}{1-\lambda_i-\lambda_j}}{(1+\frac{\lambda_i}{1-\lambda_i})(1+\frac{\lambda_j}{1-\lambda_j})}$$

- **Zero-centered Laplacian.** $p(x;\sigma) = \frac{1}{2\sigma}e^{-\frac{|x|}{\sigma}}$, $\theta = -\frac{1}{\sigma}$, $\Theta = (-\infty,0)$, $F(\theta) = \log(\frac{2}{-\theta})$.

$$\Delta_{\mathrm{Laplacian}}(\sigma_i,\sigma_j) = \log\frac{1}{2(\sigma_i+\sigma_j)}$$

- **Gaussian.** $p(x;\mu,\Sigma) =$

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right),$$

$\theta = (\theta_v,\theta_M) = (\Sigma^{-1}\mu,\Sigma^{-1})$, $\Theta = \mathbb{R}^d \times S^d_{++}$ where S^d_{++} denotes the cone of positive definite matrices of dimension $d \times d$,

$$F(\theta) = \frac{1}{2}\theta_v^T\theta_M^{-1}\theta_v - \frac{1}{2}\log|\theta_M| + \frac{d}{2}\log 2\pi.$$

The sum of natural parameters $\theta_i + \theta_j$ yields $(\Sigma_i^{-1}\mu_i + \Sigma_j^{-1}\mu_j, \Sigma_i^{-1}+\Sigma_j^{-1}) = \theta_{ij}$. This is by definition equivalent to $(\Sigma_{ij}^{-1}\mu_{ij}, \Sigma_{ij}^{-1})$ with

$$\begin{aligned}\Sigma_{ij} &= (\Sigma_i^{-1}+\Sigma_j^{-1})^{-1},\\ \mu_{ij} &= (\Sigma_i^{-1}+\Sigma_j^{-1})^{-1}(\Sigma_i^{-1}\mu_i+\Sigma_j^{-1}\mu_j)\end{aligned}$$

We express the term $\Delta_{\mathrm{Gaussian}}$ using the conventional (μ,Σ) parameterization as:

$$\begin{aligned}&\Delta_{\mathrm{Gaussian}}((\mu_i,\Sigma_i),(\mu_j,\Sigma_j)) = \frac{1}{2}\,(\\ &\mu_{ij}^T\Sigma_{ij}^{-1}\mu_{ij} - (\mu_i^T\Sigma_i^{-1}\mu_i + \mu_j^T\Sigma_j^{-1}\mu_j)\\ &-\log\frac{|\Sigma_i^{-1}+\Sigma_j^{-1}|}{|\Sigma_i^{-1}||\Sigma_j^{-1}|} - d\log 2\pi\ )\end{aligned}$$

Note that for zero-centered covariance matrices (mixtures of centered Gaussians), $\mu_{ij} = \mu_i = \mu_j = 0$ and the expression simplifies to $\Delta_{\mathrm{Gaussian}}(\Sigma_i,\Sigma_j) = -\frac{1}{2}\log\frac{|\Sigma_i^{-1}+\Sigma_j^{-1}|}{|\Sigma_i^{-1}||\Sigma_j^{-1}|} - \frac{d}{2}\log 2\pi$.

- **Wishart** [8, 6] $p(x;n,S) = \frac{|X|^{\frac{n-d-1}{2}}e^{-\frac{1}{2}\mathrm{tr}(S^{-1}X)}}{2^{\frac{nd}{2}}|S|^{\frac{n}{2}}\Gamma_d(\frac{n}{2})}$, with $S \succ 0$ the scale matrix and $n > d-1$ the number of degrees of freedom, where Γ_d is the multivariate Gamma function $\Gamma_d(x) = \pi^{d(d-1)/4}\prod_{j=1}^{d}\Gamma\left(x+(1-j)/2\right)$. $\theta = (\theta_s,\theta_M) = (\frac{n-d-1}{2}, S^{-1})$ with $\Theta = \mathbb{R}_+ \times S^d_{++}$ the cone of positive definite matrices. $F(\theta) = \frac{(2\theta_s+d+1)d}{2}\log 2 + (\theta_s + \frac{d+1}{2})\log|\theta_M| + \log\Gamma_d(\theta_s+\frac{d+1}{2})$.

3. JR and SL divergences of EFMs

The square roots of probability density functions correspond to unit vectors in the space of square integrable functions. The Hellinger distance of Eq. 7 between densities amounts to compute the L_2 norm of the difference unit vectors but does not provide a closed-form expression for mixture models. However, the squared Euclidean divergence (also called Squared Loss, or SL for short) on densities (not a metric since it does not satisfy the triangle inequality) also yields a closed-form solution for mixtures of the same exponential families: $\int(m(x)-m'(x))^2\mathrm{d}x = \int m^2(x)\mathrm{d}x - 2\int m(x)m'(x)\mathrm{d}x + \int m'^2(x)\mathrm{d}x$. That is,

$$\begin{aligned}&\sum_{i=1}^{k}\sum_{j=1}^{k} w_i w_j e^{\Delta_F(\theta_i,\theta_j)} - 2\sum_{i=1}^{k}\sum_{j=1}^{k'} w_i w'_j e^{\Delta_F(\theta_i,\theta'_j)}\\ &+\sum_{i=1}^{k'}\sum_{j=1}^{k'} w'_i w'_j e^{\Delta_F(\theta'_i,\theta'_j)}\end{aligned} \quad (8)$$

Rényi quadratic entropy H_2 is defined as $H_2(p) = -\log\int p^2(x)\mathrm{d}x$ [16], and can be computed as

$H_2(m) = -\log \sum_{i=1}^k \sum_{j=1}^k w_i w_j e^{\Delta_F(\theta_i,\theta_j)}$. Since the mixture of mixtures is a mixture, it follows that the Jensen-quadratic Rényi divergence [16]:

$$\mathrm{JR}_2(m, m') = H_2\left(\frac{m+m'}{2}\right) - \frac{H_2(m)+H_2(m')}{2}, \quad (9)$$

can also be expressed in closed-form. Similarly, the total Square Loss [11] (tSL) also admits a closed-form expression since it is defined by $\mathrm{tSL}(m : m') = \frac{\int (m(x)-m'(x))^2 \mathrm{d}x}{\sqrt{1+4\int m'(x)^2 \mathrm{d}x}}$. In particular, all those closed-form formulas apply for Wishart Mixture Models (WMMs) [6] that have been recently considered for tensor sparse coding [15] of image region covariances.

4. Concluding remarks

We derived a generic closed-form formula for the Cauchy-Schwarz/Jensen-Rényi/(total)Square Loss statistical divergences of two mixtures of the same exponential family provided that the natural parameter space Θ is a convex cone. Those properties ensure that the product of mixtures is again a mixture for those families. Our generic formula casts some light by generalizing the former *ad-hoc* expressions manually calculated for Gaussian mixtures [10], and applies to mixtures of Bernoulli [2], Wishart [6] and zero-centered Laplacian [1] distributions among others. Bypassing the stochastic approximation of KL allows us to improve by a 6-fold to 9-fold factor (million to billion) the time requires to compute a divergence between mixtures! This may prove particularly useful for information retrieval [11] and real-time applications like video tracking [15]. Closed-form expressions of divergences also allow one to design center-based clustering algorithm, where centroids are defined as minimum average divergence minimizers. For example, the total Square Loss divergence has been successfully used to design the center of a set of Gaussian mixture models, and derive an efficient shape information retrieval engine [11]. We expect similar advances for other non-Gaussian mixture models. A key open problem is to find axiomatization properties of the CS/JR/(t)SL divergences that prove theoretically their advantages other the KL divergence, and to derive rules of thumb for choosing in applications the proper dissimilarity measure in that sea of divergences.

References

[1] T. Amin, M. Zeytinoglu, and L. Guan. Application of Laplacian mixture model to image and video retrieval. *IEEE Transactions on Multimedia*, 9(7):1416–1429, 2007.

[2] Y. Amit and A. Trouvé. Generative models for labeling multi-object configurations in images. In *Toward Category-Level Object Recognition*, pages 362–381, 2006.

[3] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

[4] L. D. Brown. *Fundamentals of statistical exponential families*. Institute of Mathematical Statistics, Hayworth, CA, USA, 1986.

[5] J.-Y. Chen, J. R. Hershey, P. A. Olsen, and E. Yashchin. Accelerated Monte-Carlo for Kullback-Leibler divergence between Gaussian mixture models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4553–4556, 2008.

[6] L. Haff, P. Kim, J.-Y. Kooy, and D. Richards. Minimax estimation for mixtures of Wishart distributions. *Annals of Statistics*, 2012.

[7] J. R. Hershey and P. A. Olsen. Approximating the Kullback Leibler divergence between Gaussian mixture models. volume 4, pages 317–320, 2007.

[8] S. Hidot and C. Saint-Jean. An expectation-maximization algorithm for the Wishart mixture model: Application to movement clustering. *Pattern Recognition Letters*, 31:2318–2324, 2010.

[9] R. Jenssen, J. C. Principe, D. Erdogmus, and T. Eltoft. The Cauchy-Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *Journal of the Franklin Institute*, 343(6):614 – 629, 2006.

[10] K. Kampa, E. Hasanbelliu, and J. Principe. Closed-form Cauchy-Schwarz PDF divergence for mixture of Gaussians. In *Proc. International Joint Conference on Neural Networks (IJCNN)*, pages 2578 – 2585, 2011.

[11] M. Liu, B. C. Vemuri, S. Amari, and F. Nielsen. Shape retrieval using hierarchical total Bregman soft clustering. *Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[12] J. V. Michalowicz, J. M. Nichols, and F. Bucholtz. Calculation of differential entropy for a mixed Gaussian distribution. *Entropy*, 10(3):200–206, 2008.

[13] F. Nielsen. k-MLE: A fast algorithm for learning statistical mixture models. arXiv 1203.5181, ICASSP 2012.

[14] J. Seabra, F. Ciompi, O. Pujol, J. Mauri, P. Radeva, and J. Sanchez. Rayleigh mixture model for plaque characterization in intravascular ultrasound. *IEEE Transaction on Biomedical Engineering*, 58(5):1314–1324, 2011.

[15] R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos. Tensor sparse coding for region covariances. In *ECCV (4)*, pages 722–735, 2010.

[16] F. Wang, T. F. Syeda-Mahmood, B. C. Vemuri, D. Beymer, and A. Rangarajan. Closed-form Jensen-Rényi divergence for mixture of Gaussians and applications to group-wise shape registration. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 1, pages 648–655, 2009.