entropy

MDPI

Article

Fast Proxy Centers for the Jeffreys Centroid: The Jeffreys–Fisher–Rao Center and the Gauss–Bregman Inductive Center

Frank Nielsen

Sony Computer Science Laboratories, Tokyo 141-0022, Japan; frank.nielsen.x@gmail.com

Abstract: The symmetric Kullback–Leibler centroid, also called the Jeffreys centroid, of a set of mutually absolutely continuous probability distributions on a measure space provides a notion of centrality which has proven useful in many tasks, including information retrieval, information fusion, and clustering. However, the Jeffreys centroid is not available in closed form for sets of categorical or multivariate normal distributions, two widely used statistical models, and thus needs to be approximated numerically in practice. In this paper, we first propose the new Jeffreys–Fisher–Rao center defined as the Fisher–Rao midpoint of the sided Kullback–Leibler centroids as a plug-in replacement of the Jeffreys centroid. This Jeffreys–Fisher–Rao center admits a generic formula for uni-parameter exponential family distributions and a closed-form formula for categorical and multivariate normal distributions; it matches exactly the Jeffreys centroid for same-mean normal distributions and is experimentally observed in practice to be close to the Jeffreys centroid. Second, we define a new type of inductive center generalizing the principle of the Gauss arithmetic–geometric double sequence mean for pairs of densities of any given exponential family. This new Gauss–Bregman center is shown experimentally to approximate very well the Jeffreys centroid and is suggested to be used as a replacement for the Jeffreys centroid when the Jeffreys–Fisher–Rao center is not available in closed form. Furthermore, this inductive center always converges and matches the Jeffreys centroid for sets of same-mean normal distributions. We report on our experiments, which first demonstrate how well the closed-form formula of the Jeffreys–Fisher–Rao center for categorical distributions approximates the costly numerical Jeffreys centroid, which relies on the Lambert W function, and second show the fast convergence of the Gauss–Bregman double sequences, which can approximate closely the Jeffreys centroid when truncated to a first few iterations. Finally, we conclude this work by reinterpreting these fast proxy Jeffreys–Fisher–Rao and Gauss–Bregman centers of Jeffreys centroids under the lens of dually flat spaces in information geometry.

Keywords: Kullback–Leibler divergence; exponential family; Bregman divergence; quasi-arithmetic mean; Fisher–Rao geodesic; information geometry; Lambert W function; geometric optimization

check for updates

Citation: Nielsen, F. Fast Proxy Centers for the Jeffreys Centroid: The Jeffreys–Fisher–Rao Center and the Gauss–Bregman Inductive Center. *Entropy* **2024**, *26*, 1008. https://doi.org/10.3390/e26121008

Academic Editor: Carlo Cafaro

Received: 18 October 2024
Revised: 13 November 2024
Accepted: 21 November 2024
Published: 22 November 2024

1. Introduction

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space with sample space $\mathcal{X}$ and σ-algebra of events $\mathcal{F}$, and μ a positive measure. We consider a finite set $\{P_1, \ldots, P_n\}$ of n probability distributions all dominated by μ and weighted by a vector w belonging to the open standard simplex $\Delta_n = \{x = (x_1, \ldots, x_n) : x_1 > 0, \ldots, x_n > 0, \sum_{i=1}^n x_i = 1\} \subset \mathbb{R}^n$. Let $\mathcal{P} = \{p_1, \ldots, p_n\}$ be the Radon–Nikodym densities of $P_1, \ldots, P_n$ with respect to μ, i.e., $p_i = \frac{\mathrm{d}P_i}{\mathrm{d}\mu}$.

The Kullback–Leibler divergence (KLD) between two densities $p(x)$ and $q(x)$ is defined by $D_{\mathrm{KL}}(p : q) = \int p(x) \log \frac{p(x)}{q(x)} \,\mathrm{d}\mu(x)$. The KLD is asymmetric: $D_{\mathrm{KL}}(p : q) \neq D_{\mathrm{KL}}(q :$

p). We use the argument delimiter ':' as a notation to indicate this asymmetry. The Jeffreys divergence [1] symmetrizes the KLD as follows:

$$\begin{aligned} D_J(p,q) &= D_{\mathrm{KL}}(p:q)+D_{\mathrm{KL}}(q:p), \\ &= \int_{\mathcal{X}}(p(x)-q(x))\log\frac{p(x)}{q(x)}\mathrm{d}\mu(x). \end{aligned}$$

In general, the D-barycenter C_D of $\mathcal{P}$ with respect to a statistical dissimilarity measure $D(\cdot:\cdot)$ yields a notion of centrality C_R defined by the following optimization problem:

$$c_R=\arg\min_p\sum_{i=1}^n w_i\,D(p_i:p). \tag{1}$$

Here, the upper case letter 'R' indicates that the optimization defining the D-barycenter is carried on the right argument. When $w=(\frac{1}{n},\ldots,\frac{1}{n})$ is the uniform weight vector, the D-barycenter is called the D-centroid. We shall loosely call centroids barycenters in the remainder even when the weight vector is not uniform. Centroids with respect to information-theoretic measures have been studied in the literature.

Let us mention some examples of centroids: The entropic centroids [2] (i.e., Bregman centroids and f-divergences centroids), the Burbea–Rao and Bhattacharyya centroids [3], the α-centroids with respect to α-divergences [4], the Jensen–Shannon centroids [5], etc.

The D_J-centroid is also called the symmetric Kullback–Leibler (SKL) divergence centroid [6] in the literature. However, since there are many possible symmetrizations of the KLD [7] like the Jensen–Shannon divergence [8] or the resistor KLD [9], we prefer to use the term Jeffreys centroid instead of SKL centroid to avoid any possible ambiguity on the underlying divergence. Notice that the square root of the Jensen–Shannon divergence is a metric distance [10,11] but all powers D_J^α of Jeffreys divergence D_J for $\alpha>0$ do not yield metric distances [12].

This paper considers the Jeffreys centroids of a finite weighted set of densities $\mathcal{P}=\{p_{\theta_1},\ldots,p_{\theta_n}\}$ belonging to some prescribed exponential family [13] $\mathcal{E}$:

$$c=\arg\min_p\sum_{i=1}^n w_i\,D_J(p_{\theta_i},p). \tag{2}$$

In particular, we are interested in computing the Jeffreys centroids for sets of categorical distributions or sets of multivariate normal distributions [14].

In general, centroids are used in k means [15,16]-type clustering or hierarchical clustering (e.g., Ward criterion [17]) and information fusion tasks [18] (related to distributed model estimation [19]) among others. See Figure 1. The choice of the dissimilarity measure depends on the application at hand [20]. Clustering with respect to Jeffreys divergence/Jeffreys centroid has proven useful in many scenarios: for example, it was shown to perform experimentally better than Euclidean or square Euclidean distances for compressed histograms of gradient descriptors [21] or in fuzzy clustering [22]. Jeffreys divergence has also been used for image processing [23], including image segmentation [24], speech processing [25], and computer vision [26], just to name a few. In particular, finding weighted means of centered 3D normal distributions plays an important role in diffusion tensor imaging (DTI) for smoothing and filtering DT images [27] which consist of sets of normal distributions centered at 3D grid locations.

In general, the Jeffreys centroid is not known in closed form for exponential families [28] like the family of categorical distributions or the family of normal distributions often met in applications and thus needs to be numerically approximated in practice. The main contribution of this paper is to present and study two proxy centers as drop-in replacements of the Jeffreys centroids in applications and report the generic structural formula for generic exponential families with an explicit closed-form formula for the families of categorical and multivariate normal distributions. Namely, we define the Jeffreys–Fisher–Rao

(JFR) center (Definition 2) and the Gauss–Bregman (GB) inductive center (Definition 3) in Section 2.

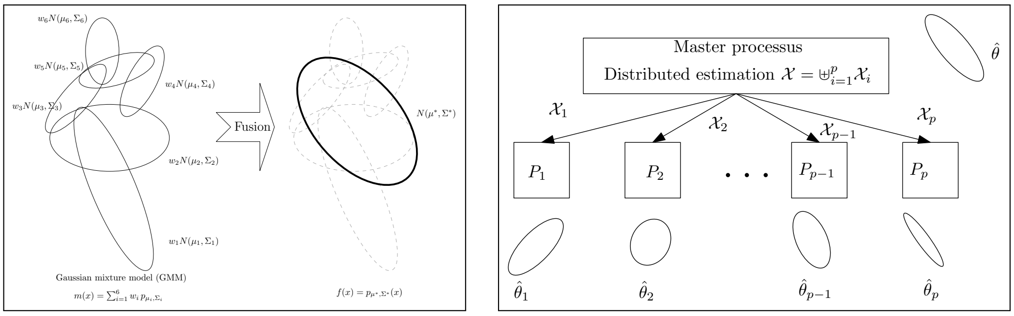

Figure 1. Application of centroids and centers in signal processing. (**Left**): information fusion and mixture model simplification, a 2D Gaussian mixture model (GMM) is simplified to a single bivariate normal distribution. (**Right**): distributed estimation, a data set is split among p processes P_is, which first estimate the statistical model parameters $\hat{\theta}_i$s. Then, the processus models are aggregated to yield a single consolidated model $\hat{\theta}$.

This paper is organized as follows: By interpreting in two different ways the closed-form formula of the Jeffreys centroids for the particular case of sets of centered multivariate normal distributions [29] (proof reported in Appendix B), we define the Gauss–Bregman (GB) centers and the Jeffreys–Fisher–Rao (JFR) centers for sets of densities belonging to an exponential family in Section 2. The Jeffreys centroid coincides with both the Gauss–Bregman inductive center and the Jeffreys–Fisher–Rao center for centered multivariate normal distributions but differ from each other in general. In Section 2.4, we study the Gauss–Bregman inductive center [30] induced by the cumulant function of an exponential family and prove the convergence under the separability condition of the generalized Gauss double sequences in the limit (Theorem 3). This Gauss–Bregman center can be easily approximated by limiting the number of iterations of a double sequence inducing it. In Section 4, we report the generic formula for Jeffreys–Fisher–Rao centers for sets of uni-order exponential families [13] and explicitly give the closed-form formula for the categorical family and the multivariate normal family. A comparison of those proxy centers with the numerical Jeffreys centroids is experimentally studied and visually illustrated with some examples. Thus, we propose to use in applications (e.g., clustering) either the fast Jeffreys–Fisher–Rao center when a closed-form formula is available for the family of distributions at hand or the Gauss–Bregman center approximation with a prescribed number of iterations as a drop-in replacement of the numerical Jeffreys centroids while keeping the Jeffreys divergence. Some experiments of the JFR and GB centers are reported for the Jeffreys centroid of categorical distributions in Section 5. Finally, we conclude this paper in Section 6 with a discussion and a generalization of our results to the more general setting of dually flat spaces of information geometry [14].

The core of this paper is followed by an Appendix section as follows: In Appendix A, we explicitly give the algorithm outlined in [31] for numerically computing the Jeffreys centroid of sets of categorical distributions. In Appendix B, we report a proof on the closed-form formula of the Jeffreys centroid for centered normal distributions [29] that motivated this paper. In Appendix C, we explain how to calculate in practice the elaborated closed-form formula for the Fisher–Rao geodesic midpoint between two multivariate normal distributions [32].

2. Proxy Centers for Jeffreys Centroids

2.1. Background on Jeffreys Centroids

A density p_θ belonging to an exponential family [13] $\mathcal{E}$ can be expressed canonically as $p_\theta(x) = \exp(\langle\theta, t(x)\rangle - F(\theta))\,\mathrm{d}\mu(x)$, where $t(x)$ is a sufficient statistic vector, $F(\theta) = \log\int\exp(\langle\theta, t(x)\rangle)\,\mathrm{d}\mu(x)$ is the log-normalizer, and θ is the natural parameter belonging to the natural parameter space Θ. We consider minimal regular exponential families [13] like the discrete family of categorical distributions (i.e., μ is the counting measure) or the continuous family of multivariate normal distributions (i.e., μ is the Lebesgue measure).

The Jeffreys centroid of categorical distributions was first studied by Veldhuis [6], who designed a numerical two-nested loops Newton-like algorithm [6]. A random variable X following a categorical distribution $\mathrm{Cat}(p)$ for a parameter $p \in \Delta_d$ in sample space $\mathcal{X} = \{\omega_1, \ldots, \omega_d\}$ is such that $\Pr(X = \omega_i) = p_i$. Categorical distributions are often used in image processing to statistically model normalized histograms with non-empty bins. The exact characterization of the Jeffreys centroid was given in [31].

We summarize the results regarding the categorical Jeffreys centroid [31] in the following theorem:

Theorem 1 (Categorical Jeffreys centroid [31])**.** *The Jeffreys centroid of a set of n categorical distributions parameterized by $\mathcal{P} = \{p_1, \ldots, p_n\} \in \Delta_d$ arranged in a matrix $P = [p_{i,j}] \in \mathbb{R}^{n \times d}$ and weighted by a vector $w = (w_1, \ldots, w_n) \in \Delta^n$ is $c(\lambda) = (c_1(\lambda), \ldots, c_d(\lambda))$ with*

$$c_j(\lambda) = \frac{a_j}{W_0\left(\frac{a_j}{g_j}\, e^{1+\lambda}\right)}, \quad \forall j \in \{1, \ldots, d\},$$

where $a_j = \sum_{i=1}^n w_i p_{i,j}$ and $g_j = \frac{\prod_{i=1}^n p_{i,j}^{w_i}}{\sum_{j=1}^d \prod_{i=1}^n p_{i,j}^{w_i}}$ are the j-th components of the weighted arithmetic and normalized geometric means, respectively; W_0 is the principal branch of the Lambert W function [33]; and $\lambda \leq 0$ is the unique real value such that $\lambda = -D_{\mathrm{KL}}(c(\lambda) : g)$.

Furthermore, a simple bisection search is reported in [31] §III.B that we convert into Algorithm A1 in Appendix A, which allows one to numerically approximate the Jeffreys centroid to arbitrary fine precision.

2.2. Jeffreys Centroids on Exponential Family Densities: Symmetrized Bregman Centroids

The Jeffreys divergence between two densities of an exponential family $\mathcal{E} = \{p_\theta(x) = \exp(\langle t(x), \theta\rangle - F(\theta)) : \theta \in \Theta\}$ with cumulant function $F(\theta)$ amounts to a symmetrized Bregman divergence [28] (SBD):

$$D_J(p_\theta, p_{\theta'}) = S_F(\theta, \theta') := \langle \theta_1 - \theta_2, \nabla F(\theta_1) - \nabla F(\theta_2)\rangle.$$

Using convex duality, we have $S_F(\theta, \theta') = S_{F^*}(\eta, \eta')$, where $\eta = \nabla F(\theta)$ and $F^*(\eta) = \langle \eta, (\nabla F)^{-1}(\eta)\rangle - F((\nabla F)^{-1}(\eta))$ is the Legendre–Fenchel convex conjugate. Thus, the Jeffreys barycenter of $\mathcal{P} = \{p_{\theta_1}, \ldots, p_{\theta_n}\}$ amounts to either a symmetrized Bregman barycenter on the natural parameters $\mathcal{P}_\theta = \{\theta_1, \ldots, \theta_n\}$ with respect to S_F or a symmetrized Bregman barycenter on the dual moment parameters $\mathcal{P}_\eta = \{\eta_1, \ldots, \eta_n\}$ with respect to S_{F^*}.

It was shown in [28] that the symmetrized Bregman barycenter θ_S of n weighted points amounts to the following minimization problem involving only the sided Bregman centroids:

$$\begin{aligned} \theta_S &:= \arg\min_{\theta \in \Theta} \sum_i w_i S_F(\theta, \theta_i), \\ &\equiv \arg\min_{\theta \in \Theta} B_F(\bar{\theta} : \theta) + B_F(\theta : \underline{\theta}), \end{aligned} \tag{3}$$

where $\bar{\theta} = \sum_i w_i \theta_i$ (right Bregman centroid) and $\underline{\theta} = (\nabla F)^{-1}(\sum_i w_i \nabla F(\theta_i))$ (left Bregman centroid). Those $\bar{\theta}$ and $\underline{\theta}$ centers are centroids [28] with respect to the Bregman divergence $B_F(\theta_1 : \theta_2) = F(\theta_1) - F(\theta_2) - \langle \theta_1 - \theta_2, \nabla F(\theta_2)\rangle$ and reverse Bregman divergence: $B_F{}^*(\theta_1 : \theta_2) := B_F(\theta_2 : \theta_1)$:

$$\begin{aligned}\bar{\theta} &= \arg\min_{\theta} \sum_i w_i B_F(\theta_i : \theta),\\ \underline{\theta} &= \arg\min_{\theta} \sum_i w_i B_F(\theta : \theta_i) = \arg\min_{\theta} \sum_i w_i {B_F}^*(\theta_i : \theta).\end{aligned}$$

In general, when $H : \mathbb{R}^m \to \mathbb{R}$ is a strictly convex differentiable real-valued function of Legendre type [34], the gradient ∇H is globally invertible (in general, the implicit inverse function theorem only locally guarantees the inverse function) and we can define a quasi-arithmetic center of a point set $\mathcal{P} = \{\theta_1, \ldots, \theta_n\}$ weighted by w as follows:

Definition 1 (Quasi-arithmetic center). *Let $H = \nabla F$ be the gradient of a strictly convex or concave differentiable real-valued function F of Legendre type. The quasi-arithmetic center $c_H(\theta_1, \ldots, \theta_n; w)$ is defined by*

$$c_H(\theta_1, \ldots, \theta_n; w) = H^{-1}\left(\sum_{i=1}^n w_i H(\theta_i)\right).$$

This definition generalizes the scalar quasi-arithmetic means [35] for univariate functions h, which are continuous and strictly monotone. Quasi-arithmetic means (QAMs) are also called f means or Kolmogorov–Nagumo means. Let $m_{\nabla F}(\theta_1, \theta_2) = c_{\nabla F}(\theta_1, \theta_2; \frac{1}{2}, \frac{1}{2})$. Notice that $A(\theta_1, \theta_2) = \nabla F^*(m_{\nabla F^*}(\eta_1, \eta_2))$ and $A(\eta_1, \eta_2) = \nabla F^{(} m_{\nabla F}(\theta_1, \theta_2))$. That is, the arithmetic mean in a primal representation amounts to a QAM in the dual representation.

Thus, we can solve for θ_S by setting the gradient of $L(\theta) = B_F(\bar{\theta} : \theta) + B_F(\theta : \underline{\theta})$ to zero. In general, no closed-form formula is known for the symmetrized Bregman centroids, and a numerical approximation method was reported in [28]. To circumvent the lack of a closed-form formula of symmetrized Bregman centroids for clustering, Nock et al. [36] proposed a mixed Bregman clustering where each cluster has two representative dual Bregman centroids $\bar{\theta} = \sum_i w_i \theta_i$ (right Bregman centroid) and $\underline{\theta} = (\nabla F)^{-1}(\sum_i w_i \nabla F(\theta_i))$ (left Bregman centroid), and the dissimilarity measure is a mixed Bregman divergence defined by

$$\Delta_F(\theta_1 : \theta : \theta_2) := \frac{1}{2} B_F(\theta_1 : \theta) + \frac{1}{2} B_F(\theta : \theta_2).$$

Notice that minimizing Equation (3) amounts to minimizing the mixed Bregman divergence:

$$\min_{\theta} \Delta_F(\bar{\theta} : \theta : \underline{\theta}).$$

By using the dual parameterization $\eta = \nabla F(\theta)$ (with dual domain $H = \{\nabla F(\theta) : \theta \in \Theta\}$) and the dual Bregman divergence $B_{F^*}(\eta_1 : \eta_2) = F^*(\eta_1) - F^*(\eta_2) - \langle \eta_1 - \eta_2, \nabla F\rangle^*(\eta_1) = B_F(\theta_2 : \eta_1)$, we have

$$\begin{aligned}\theta_S &:= \arg\min_{\theta \in \Theta} \sum_i w_i S_F(\theta, \theta_i),\\ \eta_S &= \arg\min_{\eta \in H} \sum_i w_i S_{F^*}(\eta, \eta_i),\\ &\equiv \arg\min_{\eta \in H} B_{F^*}(\eta : \nabla F(\bar{\theta})) + B_{F^*}(\nabla F(\underline{\theta}) : \eta). \qquad (4)\end{aligned}$$

Since $\nabla F(\bar{\theta}) = (\nabla F^*)^{-1}(\sum_i w_i \nabla F^*(\eta_i)) = \underline{\eta}$ and $\nabla F(\underline{\theta}) = \nabla F((\nabla F)^{-1} \sum_i w_i \eta_i) = \bar{\eta}$, we obtain the dual equivalent optimization problem:

$$\theta_S = \nabla F^*(\eta_S) = \arg\min_{\theta \in \Theta} B_F(\bar{\theta} : \theta) + B_F(\theta : \underline{\theta}),$$

or

$$\eta_S = \nabla F(\theta_S) = \arg\min_{\eta \in H} B_{F^*}(\bar{\eta} : \theta) + B_{F^*}(\eta : \underline{\eta}).$$

However, a remarkable special case is the family of multivariate normal distributions centered at the origin for which the Jeffreys centroid was reported in closed form in [29]. Let $\mathcal{N}_0 = \{p_\Sigma : \Sigma \in \mathrm{Sym}^{++}(\mathbb{R}, d)\}$ be the exponential family with sufficient statistics $t(x) = -\frac{1}{2}(x, xx^\top)$, natural parameter $\theta = \Sigma^{-1}$ (the precision matrix) where the covariance matrix belongs to the cone $\mathrm{Sym}^{++}(\mathbb{R}, d)$ of symmetric positive-definite matrices, inner product $\langle X, Y\rangle = \mathrm{tr}(XY)$, and $F(\theta) = -\frac{1}{2}\log\det(\theta)$. In that case, the Jeffreys divergence amounts to a symmetrized Bregman log-det (ld) divergence between the corresponding natural parameters:

$$D_J(p_\Sigma, p_{\Sigma'}) = \frac{1}{2}\,\mathrm{tr}\Big(\big(\Sigma'^{-1} - \Sigma^{-1}\big)(\Sigma - \Sigma')\Big) =: \frac{1}{2} S_{\mathrm{ld}}(\Sigma^{-1}, \Sigma'^{-1}).$$

Using the standard covariance matrix parameterization Σ, we can further express the Jeffreys divergence between two multivariate normal distributions p_Σ and $p_{\Sigma'}$ as

$$D_J(p_\Sigma, p_{\Sigma'}) = \sum_{i=1}^{d}\left(\sqrt{\lambda_i} - \frac{1}{\sqrt{\lambda_i}}\right)^2,$$

where λ_is are the eigenvalues of $\Sigma^{-1}\Sigma'$. The symmetrized log-det divergence S_{ld} is also called the symmetrized Stein loss [37,38]. When $d = 1$, this divergence is the symmetrized Itakura–Saito divergence also called the COSH distance [28]. The Jeffreys centroid can be characterized using the Fisher–Rao geometry [39] of $\mathcal{N}_0$ as the Fisher–Rao geodesic midpoint of the sided Kullback–Leibler centroids as follows:

Theorem 2 ([29])**.** *The Jeffreys centroid C of a set of n centered multivariate normal distributions $\mathcal{P} = \{p_{\Sigma_1}, \ldots, p_{\Sigma_n}\}$ weighted with $w_i \in \Delta_n$ amounts to the symmetrized log-det Bregman centroid for the corresponding weighted set of positive-definite precision matrices $\mathcal{P}_\theta = \{P_1 = \Sigma_1^{-1}, \ldots, P_n = \Sigma_n^{-1}\}$. The symmetrized log-det Bregman barycenter C is the Riemannian geodesic midpoint $A\#H$ of the arithmetic barycenter $A = \sum_{i=1}^n w_i P_i$ and harmonic barycenter $H = \left(\sum_{i=1}^n w_i P_i^{-1}\right)^{-1}$ where $X\#Y := X^{\frac{1}{2}}\left(X^{-\frac{1}{2}}\, Y\, X^{-\frac{1}{2}}\right)^{\frac{1}{2}} X^{\frac{1}{2}}$ is the matrix geometric mean [40] $G(X, Y) = X\#Y$:*

$$C = (\sum_{i=1}^{n} w_i P_i)\#\left(\sum_{i=1}^{n} w_i P_i^{-1}\right)^{-1}. \tag{5}$$

Since the proof of this result mentioned in [29] was omitted in [29], we report a proof involving matrix analysis in full detail in Appendix B.

Next, we shall define two types of centers for sets of densities of a prescribed exponential family based on two different interpretations of Equation (5). We call them centers and not centroids because those points are defined by a generic structural formula instead of solutions of minimization problems of average divergences of Equation (1).

2.3. The Jeffreys–Fisher–Rao Center

Since an exponential family $\mathcal{E} = \{p_\theta(x)\}$ induces the Riemannian manifold $(\mathcal{M}, g)$ with the Fisher metric g expressed in the θ-parameterization by the Fisher information matrix $\nabla^2 F(\theta)$ and Fisher–Rao geodesics $\gamma(p, q, t)$ defined with respect to the Levi-Civita connection $\bar{\nabla}$ (induced by g), we shall define the Jeffreys–Fisher–Rao center on $\mathcal{M}$ using the Fisher–Rao geodesics as follows:

Definition 2 (Jeffreys–Fisher–Rao (JFR) center)**.** *The Jeffreys–Fisher–Rao center θ_{JFR} of a set $\{p_{\theta_1}, \ldots, p_{\theta_n}\}$ of weighted densities by $w \in \Delta_n$ is defined as the Fisher–Rao midpoint of the sided Kullback–Leibler centroids $\bar{\theta} = \sum_i w_i \theta_i$ and $\underline{\theta} = (\nabla F)^{-1}(\sum_i w_i \nabla F(\theta_i))$:*

$$\theta_{\mathrm{JFR}} = \bar{\theta}\#\underline{\theta}, \tag{6}$$

where $p\#q = \gamma\left(p, q, \frac{1}{2}\right)$.

Equation (6) is a generalization of Equation (5); therefore, the JFR center matches the Jeffreys centroid for same-mean multivariate normal distributions (Theorem 2).

Let $\mathcal{P}_\theta = \{\theta_1, \ldots, \theta_n\}$ and $\mathcal{P}_\eta = \{\eta_1, \ldots, \eta_n\}$, where $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$. Denote by $\mathrm{JFR}_F(\mathcal{P}_\theta; w)$ the JFR center of θ-coordinate θ_{JFR}. Then, $\mathrm{JFR}_{F^*}(\mathcal{P}_\eta; w)$ = $\nabla F(\theta_{\mathrm{JFR}}) := \eta_{\mathrm{JFR}}$.

2.4. Gauss–Bregman Inductive Center

Another remarkable property of the Jeffreys centroid for a set $\{p_{\mu,\Sigma_1}, \ldots, p_{\mu,\Sigma_n}\}$ of same-mean multivariate normal distributions weighted by $w \in \Delta_n$ with arithmetic and harmonic means $A = \sum_{i=1}^n w_i \Sigma_i^{-1}$ and $H = (\sum_{i=1}^n w_i \Sigma_i)^{-1}$ on the precision matrices $\Sigma_1^{-1}, \ldots, \Sigma_n^{-1}$, respectively, is that we have the following invariance of the Jeffreys centroid (see Lemma 17.4.4 of [29]):

$$G(A, H) = G\left(\frac{A+H}{2}, 2\,(A^{-1} + H^{-1})^{-1}\right). \tag{7}$$

Nakamura [41] defined the following double sequence scheme converging to the matrix geometry mean $G(P, Q)$ for any two symmetric positive-definite matrices P and Q:

$$\begin{aligned} P_{t+1} &= A(P_t, Q_t) := \frac{P_t + Q_t}{2}, \\ Q_{t+1} &= H(P_t, Q_t) := 2\,(P_t^{-1} + Q_t^{-1})^{-1}, \end{aligned}$$

initialized with $P_0 = P$ and $Q_0 = Q$. We have $\lim_{t\to\infty} P_t = \lim_{t\to\infty} Q_t = P\#Q = G(P, Q)$. Let $P_\infty = \lim_{t\to\infty} P_t$ and $Q_\infty = \lim_{t\to\infty} Q_t$. That is, the geometric matrix mean can be obtained as the limits of a double sequence of means. We can thus approximate $G(P, Q)$ by stopping the double sequence after T iterations to obtain

$$G^{(T)}(P, Q) = A(P_T, Q_T) \approx G(P, Q).$$

Notice that we can recover those iterations from the invariance property of Equation (7): Indeed, we have

$$G(P_0, Q_0) = G(\underbrace{A(P_0, Q_0)}_{=:P_1}, \underbrace{H(P_0, Q_0)}_{=:Q_1}) = G(\underbrace{A(P_1, Q_1)}_{=:P_2}, \underbrace{H(P_1, Q_1)}_{=:Q_2}) = \ldots, \tag{8}$$

and $\|P_t - Q_t\| = \sqrt{\mathrm{tr}((P_t - Q_t)(P_t - Q_t))}$ decreases [41] as the number of iterations t increases. Thus, by induction, $G(P_0, Q_0) = G(P_\infty, Q_\infty)$ with $P_\infty = Q_\infty$. Since $G(X, X) = X$ (means are reflexive), it follows that $G(P_0, Q_0) = P_\infty = Q_\infty$. It is proved in [41] that the convergence rate of the sequence of double iterations is quadratic. This type of mean has been called an inductive mean [30,42] (or compound mean [43]) and originated from the Gauss arithmetic–geometric mean [44].

Our second interpretation of the geometric matrix mean of Equation (5) is to consider it as an inductive mean [30] and to generalize this double sequence process to pairs/sets of densities of an exponential family as follows:

Definition 3 (Gauss–Bregman $(A, \nabla F)$ center)**.** *Let $\mathcal{P} = \{p_{\theta_1}, \ldots, p_{\theta_n}\}$ be a set of n distributions of an exponential family with the cumulant function $F(\theta)$ weighted by a vector $w \in \Delta_n$. Then, the Gauss–Bregman inductive center θ_{GB} is defined as the limit of the double sequence:*

$$\begin{aligned}
\bar{\theta}_{t+1} &= A(\bar{\theta}_t, \underline{\theta}_t) := \frac{\bar{\theta}_t + \underline{\theta}_t}{2}, \\
\underline{\theta}_{t+1} &= m_{\nabla F}(\bar{\theta}_t, \underline{\theta}_t) := (\nabla F)^{-1}\left(\frac{\nabla F(\bar{\theta}_t) + \nabla F(\underline{\theta}_t)}{2}\right),
\end{aligned}$$

initialized with $\bar{\theta}_0 = \bar{\theta} = \sum_{i=1}^n w_i \theta_i$ *(right Bregman centroid) and* $\underline{\theta}_0 = \underline{\theta} = \nabla F^{-1}(\sum_{i=1}^n w_i \nabla F(\theta_i))$ *(left Bregman centroid). That is, we have*

$$\theta_{\mathrm{GB}} = \lim_{t\to\infty} \bar{\theta}_t = \lim_{t\to\infty} \underline{\theta}_t. \tag{9}$$

Let $\theta_{\mathrm{GB}} = \mathrm{GB}_F(\bar{\theta}, \underline{\theta})$. Then, we have $\eta_{\mathrm{GB}} = \mathrm{GB}_{F^*}(\bar{\eta}, \bar{\theta}) = \nabla F(\theta_{\mathrm{GB}})$. The Gauss–Bregman center c_{GB} has θ-coordinates $\theta_G B$ and η-coordinates η_{GB}.

Algorithm 1 describes the approximation of the Gauss–Bregman inductive center by stopping the double sequence when the iterated centers are close enough to each other. We shall prove the matching convergence of those $\bar{\theta}_t$ and $\underline{\theta}_t$ sequences under separability conditions in Section 2.4.

Algorithm 1: Gauss–Bregman inductive center.

```
Input: A set $\mathcal{P} = \{p_{\theta_1}, \ldots, p_{\theta_n}\}$ of weighted densities with $w \in \Delta_n$ of an
       exponential family with cumulant function $F(\theta)$, natural parameters $\theta_i$'s lie
       in an inner product space $(\Theta, \langle\cdot,\cdot\rangle)$.
Input: The distance is defined as $\|\theta - \theta'\| = \sqrt{\langle \theta - \theta', \theta - \theta' \rangle}$
Input: A precision parameter $\epsilon > 0$
Output: A numerical approximation of the symmetrized Bregman centroid
/* Arithmetic weighted mean on natural parameters */
$\bar{\theta}_0 = \sum_{i=1}^n w_i \theta_i$ ;
/* Dual weighted mean */
$\underline{\theta}_0 = \nabla F^{-1}(\sum_{i=1}^n w_i \nabla F(\theta_i))$ ;
$t \leftarrow 0$;
/* Iterate until close to convergence */
while $|\bar{\theta}_t - \underline{\theta}_t| > \epsilon$ do
    $\bar{\theta}_{t+1} = \frac{\bar{\theta}_t + \underline{\theta}_t}{2}$ ;
    $\underline{\theta}_{t+1} = \nabla F^{-1}\left(\frac{\nabla F(\bar{\theta}_t) + \nabla F(\underline{\theta}_t)}{2}\right)$ ;
    $t \leftarrow t + 1$;
end
return $\bar{\theta}_{t-1}$;
```

For example, the Gauss–Bregman center of two categorical distributions $p = (p_1, \ldots, p_d)$ and $p' = (p'_1, \ldots, p'_d)$ on a sample space $\mathcal{X}$ of d elements is obtained for the cumulant function $F(\theta) = \log(1 + \sum_{i=1}^{d-1} e^{\theta_i})$ with gradient $\nabla F(\theta) = \left[\eta_i = \frac{e^{\theta_i}}{1 + \sum_{j=1}^{d-1} e^{\theta_j}}\right]_i$ where $\theta = (\theta_1 = \log \frac{p_1}{p_d}, \ldots, \theta_{d-1} = \log \frac{p_{d-1}}{p_d})$ is the natural parameter. The reciprocal gradient is $(\nabla F)^{-1}(\eta) = \left[\log \frac{\eta_i}{1 - \sum_{j=1}^{d-1} \eta_j}\right]_i$.

We may also compute the Gauss–Bregman center of two categorical distributions $\mathrm{Cat}(p)$ and $\mathrm{Cat}(p')$ using iterations of arithmetic means a_t and geometric normalized means g_t:

$$
\begin{aligned}
a_{t+1}^i &= A(a_t^i, g_t^i) := \frac{a_t^i + g_t^i}{2}, \quad \forall i \in \{1, \ldots, d\} \\
u_{t+1}^i &= \sqrt{a_t^i g_t^i}, \quad \forall i \in \{1, \ldots, d\}, \\
g_{t+1}^i &= \frac{u_{t+1}^i}{\sum_{j=1}^d u_{t+1}^j}, \quad \forall i \in \{1, \ldots, d\},
\end{aligned}
$$

where the u_ts are unnormalized geometric means and the g_t represents normalized geometric means. We initialize the sequence with $a_0 = p$ and $g_0 = p'$, and the Gauss–Bregman center is obtained in the limit $m_{\mathrm{GB}}^{\mathrm{Cat}}(p, p') = \lim_{t \to \infty} a_t = \lim_{t \to \infty} g_t$. See Algorithm 2.

The Jeffreys centroid of a set of centered multivariate normal distributions is the Gauss–Bregman center obtained for the generator $F(\theta) = -\frac{1}{2} \log \det(\theta)$, the cumulant function of the exponential family of centered normal distributions.

Algorithm 2: Gauss–Bregman inductive center for sets of categorical distributions.

```
Input: A set of weighted categorical distributions: $\mathcal{P}^w = \{p_1, \ldots, p_n\}$ with
       $w \in \Delta_n$ and $p_i \in \Delta_d$. Let $p_{i,j}$ denote the $j$-th component of $p_i$.
Input: A precision parameter $\epsilon > 0$
Input: Distance is chosen as total variation $\frac{1}{2}\|\cdot\|_1$
Output: A numerical approximation of the SKL centroid/Jeffreys centroid $c$
/* Arithmetic weighted mean (normalized)                                  */
$a_0^j = \sum_{i=1}^n w_i p_{i,j}$ for $i \in \{1, \ldots, d\}$ for $j \in \{1, \ldots, d\}$ ;
/* Normalized geometric weighted mean                                     */
$g_0^j = \frac{\prod_{i=1}^n p_{i,j}^{w_i}}{\sum_{j=1}^d \prod_{i=1}^n p_{i,j}^{w_i}}$ for $j \in \{1, \ldots, d\}$ ;
$t \leftarrow 0$;
/* Iterate until close to convergence                                     */
while $\|a_t - g_t\|_1 > 2\epsilon$ do
    /* Arithmetic mean                                                    */
    $a_{t+1}^i = \frac{a_t^i + g_t^i}{2}, \quad \forall i \in \{1, \ldots, d\}$
    /* Non-normalized geometric mean                                      */
    $u_{t+1}^i = \sqrt{a_t^i g_t^i}, \quad \forall i \in \{1, \ldots, d\}$
    /* Normalized geometric mean                                          */
    $g_{t+1}^i = \frac{u_{t+1}^i}{\sum_{j=1}^d u_{t+1}^j}, \quad \forall i \in \{1, \ldots, d\}$
    $t \leftarrow t + 1$;
end
return $a^{(t-1)}$;
```

Figure 2 displays the arithmetic and normalized geometric and numerical Jeffreys, Jeffreys–Fisher–Rao, and Gauss–Bregman centroids/centers for a set of 32 trinomial distributions. We may consider normalized intensity histograms of images (modeled as multinomials with one trial) quantized with $d = 256$ bins; that is, a normalized histogram with d bins is interpreted as a point in Δ_d and visualized as a polyline with $d - 1$ line segments. Figure 3 (left) displays the various centroids and centers obtained for an input set consisting of two histograms (the commonly used `Barbara` and `Lena` images, which have been used in [31]). Notice that the JFR center (purple) and GB center (yellow) are close to the numerical Jeffreys centroid (green). We also provide a close-up window in Figure 3 (right).

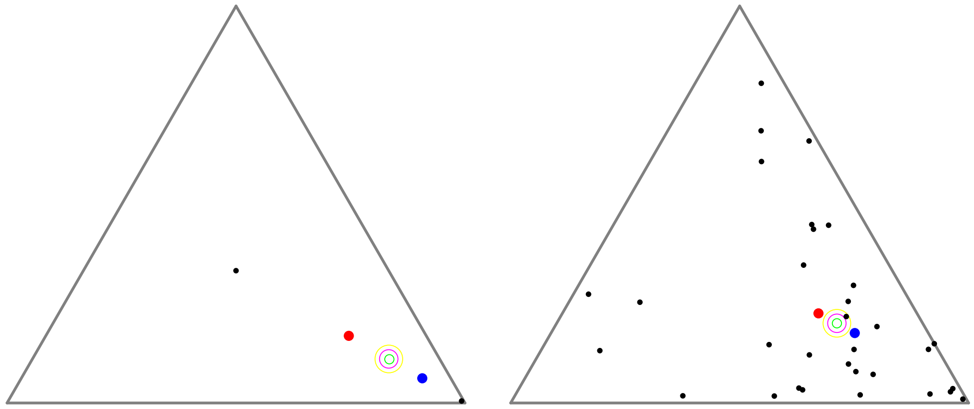

Figure 2. Visualizing the arithmetic and normalized geometric and numerical Jeffreys, Jeffreys–Fisher–Rao, and Gauss–Bregman centroids/centers in red, blue, green, purple, and yellow, respectively. (**Left**): input set consists of $n = 2$ trinomial distributions (black) with parameters chosen randomly. (**Right**): input set consists of $n = 32$ trinomial distributions (black) with parameters $(\frac{1}{2}, \frac{1}{2})$ and $(0.99, 0.005, 0.005)$. The numerical Jeffreys centroid (green) is time consuming to calculate using the Lambert W function. However, the Jeffreys centroid can be well approximated by either the Jeffreys–Fisher–Rao center (purple) or the inductive Gauss–Bregman center (yellow). Point centers are visualized with different radii in order to distinguish them easily.

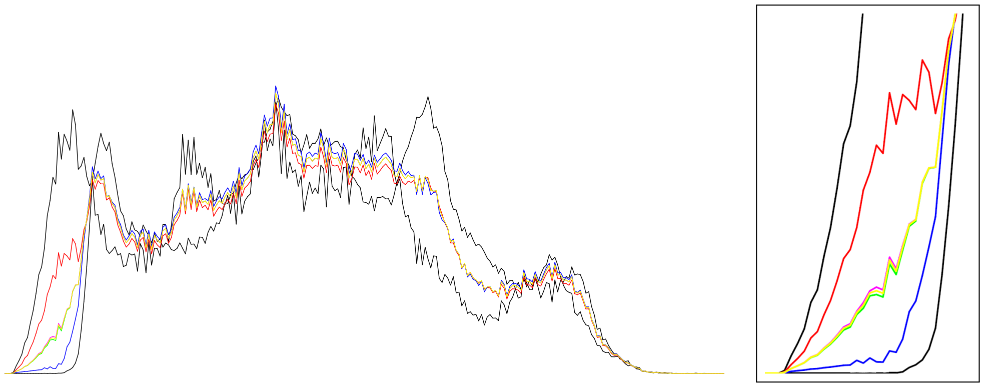

Figure 3. (**Left**): Displaying the arithmetic and normalized geometric and numerical Jeffreys, Jeffreys–Fisher–Rao, and Gauss–Bregman centroids/centers in red, blue, green, purple, and yellow, respectively. Input sets are two normalized histograms with $d = 256$ bins plotted as polylines with 255 line segments (black). Observe that the Jeffreys–Fisher–Rao center (purple) and Gauss–Bregman center (yellow) approximates the Jeffreys centroid (green) well, which is more computationally expensive to calculate. (**Right**): close-up window on the first left bins of normalized histograms.

Notice that we can experimentally check the quality of the approximation of the Gauss–Bregman center to the Jeffreys centroid by defining the symmetrized Bregman centroid energy:

$$E_F(\theta) := \langle \theta - \bar{\theta}, \nabla F(\theta) \rangle - \langle \theta, \nabla F(\bar{\theta}) \rangle,$$

and checking that $\nabla E_F(\theta)$:

$$\forall i, \qquad \partial_i \left(\sum_{i=1}^{d} (\theta_i - \bar{\theta}_i) \partial_i F(\theta) - \theta_i \partial_i F(\bar{\theta}) \right) = 0, \tag{10}$$

$$\partial_i F(\theta) + (\theta_i - \bar{\theta}_i) \partial_i^2 F(\theta) - \partial_i F(\bar{\theta}) + \left(\sum_{j \neq i} (\theta_j - \bar{\theta}_j) \partial_i \partial_j F(\theta) - \partial_i \theta_j \partial_j F(\bar{\theta}) \right) = 0 \tag{11}$$

is close to zero, where $\partial_l := \frac{\partial}{\partial \theta_l}$.

Next, we study these two new types of centers and how well they approximate the Jeffreys centroid.

3. Gauss–Bregman Inductive Centers: Convergence Analysis and Properties

Let $F(\theta)$ be a strictly convex and differentiable real-valued function of Legendre type [45] defined on an open parameter space Θ. Then, the gradient map $\theta \mapsto \eta(\theta) = \nabla F(\theta)$ is a bijection with the reciprocal inverse function] $\eta \mapsto \theta(\eta) = \nabla F^*(\eta) = (\nabla F)^{-1}(\eta)$ where $F^*(\eta) = \langle \eta, \nabla F^{-1}(\eta)\rangle - F(\nabla F^{-1}(\eta))$ is the Legendre–Fenchel convex conjugate. For example, we may consider the cumulant functions of regular exponential families.

We define the Gauss–Bregman center θ_{GB} of a set $\{\theta_1, \ldots, \theta_n\}$ weighted by $w \in \Delta_n$ as the limits of the sequences $\bar{\theta}_1, \ldots$ and $\underline{\theta}_1, \ldots$ defined by

$$\bar{\theta}_{t+1} = A(\bar{\theta}_t, \underline{\theta}_t) := \frac{\bar{\theta}_t + \underline{\theta}_t}{2}, \quad (12)$$

$$\underline{\theta}_{t+1} = m_{\nabla F}(\bar{\theta}_t, \underline{\theta}_t) := (\nabla F)^{-1}\left(\frac{\nabla F(\bar{\theta}_t) + \nabla F(\underline{\theta}_t)}{2}\right), \quad (13)$$

initialized with $\bar{\theta}_0 = \bar{\theta} = \sum_{i=1}^n w_i \theta_i$ and $\underline{\theta}_0 = \underline{\theta} = \nabla F^{-1}(\sum_{i=1}^n w_i \nabla F(\theta_i))$. That is, we have

$$\theta_{\mathrm{GB}} = \lim_{n\to\infty} \bar{\theta}_t = \lim_{n\to\infty} \underline{\theta}_t.$$

Such a center has been called an inductive mean by Sturm [30]. See [42] for an overview of inductive means. Figure 4 geometrically illustrates the double sequence iterations converging to the Gauss–Bregman mean.

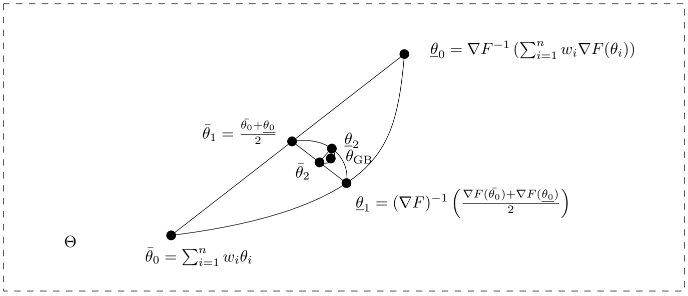
$\underline{\theta}_0 = \nabla F^{-1}\left(\sum_{i=1}^n w_i \nabla F(\theta_i)\right)$
$\bar{\theta}_1 = \frac{\bar{\theta}_0 + \underline{\theta}_0}{2}$
$\underline{\theta}_2$
θ_{GB}
$\bar{\theta}_2$
$\underline{\theta}_1 = (\nabla F)^{-1}\left(\frac{\nabla F(\bar{\theta}_0)+\nabla F(\underline{\theta}_0)}{2}\right)$
Θ
$\bar{\theta}_0 = \sum_{i=1}^n w_i \theta_i$

Figure 4. Geometric illustration of the double sequence inducing a Gauss–Bregman center in the limit.

Theorem 3. *The Gauss–Bregman* $(A, \nabla F)$ *center with respect to a Legendre type function* $F(\theta)$ *is well defined (i.e., the double sequence converges) for separable Bregman generators.*

Proof. We need to prove the convergence of $\{\bar{\theta}_t\}$ and $\{\underline{\theta}_t\}$ to the same finite limit. When $F(\theta)$ is univariate, the convergence of the inductive centers was reported in [43]. We need to prove that the double iterations of Equation (13) and Equation (13) converge.

Let us consider the following cases:

1. When the dimension is one, the quasi-arithmetic mean $m_{f'}$ for f, a strictly convex and differentiable function, lies between the minimal and maximal argument (i.e., this is the definition of a strict mean):

   $$\min\{\theta_1, \theta_2\} \leq m_{f'}(\theta_1, \theta_2) \leq \max\{\theta_1, \theta_2\}.$$

   Thus, we have

   $$|\bar{\theta}_{t+1} - \underline{\theta}_{t+1}| \leq \frac{1}{2}|\bar{\theta}_t - \underline{\theta}_t|,$$

   and it follows that $|\bar{\theta}_{t+1} - \underline{\theta}_{t+1}| \leq \frac{1}{2^t}|\bar{\theta}_0 - \underline{\theta}_0|$. Thus, we have quadratic convergence of scalar (A, f') means. See Figure 5.

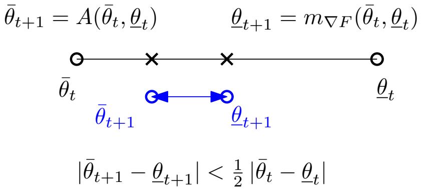
$\bar{\theta}_{t+1} = A(\bar{\theta}_t, \underline{\theta}_t)$
$\underline{\theta}_{t+1} = m_{\nabla F}(\bar{\theta}_t, \underline{\theta}_t)$
$\bar{\theta}_t$
$\underline{\theta}_t$
$\bar{\theta}_{t+1}$
$\underline{\theta}_{t+1}$
$|\bar{\theta}_{t+1} - \underline{\theta}_{t+1}| < \frac{1}{2}|\bar{\theta}_t - \underline{\theta}_t|$

Figure 5. Illustration of the double sequence convergence for scalar Gauss–Bregman $(A, m_{\nabla F})$ mean.

2. When $F(\theta)$ is multivariate and separable, i.e., $F(\theta) = \sum_{i=1}^d f_i(\theta^i)$ where $\theta = (\theta^1, \ldots, \theta^d)$ are the components of $\theta \in \mathbb{R}^d$ and the f_is are scalar strictly convex and differentiable functions, we can apply case 1 dimension-wise to obtain the quadratic convergence.
3. Otherwise, we consider the multivariate quasi-arithmetic center $m_{\nabla F}(\theta, \theta')$ with the uniform weight vector $w = (\frac{1}{2}, \frac{1}{2})$. One problem we face is that the quasi-arithmetic center $m_{\nabla F}(\theta, \theta')$ for $\theta \neq \theta'$ may lie outside the open bounding box of $\mathbb{R}^d$ with diagonal corners θ and θ':

$$\theta_m = (\min\{\theta^1, \theta'^1\}, \ldots, \min\{\theta^d, \theta'^d\}), \quad \theta_M = (\max\{\theta^1, \theta'^1\}, \ldots, \max\{\theta^d, \theta'^d\}).$$

   Indeed, in the 2D case, we may consider $\theta = (x, y)$ and $\theta' = (x', y)$. Clearly, the open bounding box is empty, and the midpoint $m_{\nabla F}(\theta, \theta')$ lies outside this box. Yet, we are interested in the convergence rate when $\theta' \approx \theta$.

   In general, we shall measure the difference between two iterations by the squared norm distance induced by the inner product:

$$\|A(\theta, \theta') - m_{\nabla F}(\theta, \theta')\|^2 = \langle A(\theta, \theta') - m_{\nabla F}(\theta, \theta'), A(\theta, \theta') - m_{\nabla F}(\theta, \theta') \rangle.$$

□

Let $m_F^{\mathrm{GB}}(\theta_1, \theta_2)$ denote the Gauss–Bregman center of θ_1 and θ_2, $A(\theta_1, \theta_2) = \frac{\theta_1 + \theta_2}{2}$ the arithmetic mean, and $m_{\nabla F}(\theta_1, \theta_2) = (\nabla F)^{-1}\left(\frac{\nabla F(\theta_1) + \nabla F(\theta_2)}{2}\right)$ the quasi-arithmetic center.

By construction, the Gauss–Bregman center enjoys the following invariance property generalizing Lemma 17.4.4 of [29] in the case of the log det generator:

Property 1. *We have* $m_F^{\mathrm{GB}}(\theta_1, \theta_2) = m_F^{\mathrm{GB}}(A(\theta_1, \theta_2), m_{\nabla F}(\theta_1, \theta_2))$.

Proof. Similar to the cascaded inequalities of Equation (8), we have

$$m_F^{\mathrm{GB}}(\theta_1, \theta_2) = m_{\mathrm{GB}}^F(\underbrace{A(\theta_1, \theta_2)}_{=:\theta_1^{(1)}}, \underbrace{m_{\nabla F}(\theta_1, \theta_2)}_{=\theta_2^{(1)}}) = \ldots \tag{14}$$

In the limit $t \to \infty$, we have $m_F^{\mathrm{GB}}(\theta_1, \theta_2) = m_F^{\mathrm{GB}}(\theta_1^{(\infty)}, \theta_2^{(\infty)}) = m_F^{\mathrm{GB}}(\theta_1^{(\infty-1)}, \theta_2^{(\infty-1)}) = \ldots$ Since $\infty - 1 = \infty$, we obtain the desired invariance property:

$$m_F^{\mathrm{GB}}(\theta_1, \theta_2) = m_F^{\mathrm{GB}}(A(\theta_1, \theta_2), m_{\nabla F}(\theta_1, \theta_2)).$$

□

Note that when $F(\theta)$ is univariate, the Gauss–Bregman mean $m_F^{\mathrm{GB}}(\theta_1, \theta_2)$ converges at a quadratic rate [43]. In particular, when $F(\theta) = -\log\theta$ (Burg negentropy), we have $F'(\theta) = -\frac{1}{\theta}$ ($m_{F'}$ is the harmonic mean) and the Gauss–Bregman mean is the arithmetic–harmonic mean (AHM) which converges to the geometric mean, a simple closed-form formula. Notice that the geometric mean $g = \sqrt{xy}$ of two scalars $x > 0$ and $y > 0$ can be expressed using the arithmetic mean $a = \frac{x+y}{2}$ and the harmonic mean $h = \frac{2xy}{x+y}$: $g = \sqrt{ah}$. But when $F(\theta) = \theta\log\theta - \theta$ (Shannon negentropy), the Gauss–Bregman mean $m_F^{\mathrm{GB}}(\theta_1, \theta_2)$ coincides with the Gauss arithmetic–geometric mean [44] (AGM) since $F'(\theta) = \log\theta$ and $m_{F'}(\theta_1, \theta_2) = \sqrt{\theta_1\theta_2}$, the geometric mean. Thus, $m_F^{\mathrm{GB}}(\theta_1, \theta_2)$ is related to the elliptic integral K of the first type [44]: there is no closed-form formula for the AGM in terms of elementary functions as this induced mean is related to the complete elliptic integral of the first kind $K(\cdot)$:

$$\mathrm{AGM}(x,y) = \frac{\pi}{4}\frac{x+y}{K\left(\frac{x-y}{x+y}\right)}, \tag{15}$$

where $K(u) = \int_0^{\frac{\pi}{2}} \frac{\mathrm{d}\theta}{\sqrt{1-u^2\sin^2(\theta)}}$ is the elliptic integral. Thus, it is difficult, in general, to report a closed-form formula for the inductive Gauss–Bregman means even for univariate generators $F(\theta)$.

The Jeffreys centroid of $x > 0$ and $y > 0$ with respect to the scalar Jeffreys divergence $D_J(p,q) = (p-q)\log\frac{p}{q}$ admits a closed-form solution [31]:

$$c = \frac{a}{W_0\left(\frac{a}{g}e\right)} \tag{16}$$

where $a = \frac{x+y}{2}$ and $g = \sqrt{xy}$ and W_0 is the principal branch of the Lambert W function [33]. This example shows that the Gauss–Bregman center does not coincide with the Jeffreys centroid in general (e.g., compare Equation (15) with Equation (16)).

4. Jeffreys–Fisher–Rao Centers: Generic Structural Formula and Some Closed-Form Formula

4.1. Jeffreys–Fisher–Rao Center for Uni-Parametric Statistical Models

Consider a set $\mathcal{P} = \{p_{\theta_1}, \ldots, p_{\theta_n}\}$ of n parametric distributions where $\theta \in \Theta \subset \mathbb{R}$ is a scalar parameter. Let $w = (w_1, \ldots, w_n) \in \Delta_n$ be a weight vector on $\mathcal{P}$ such that the weight of p_{θ_i} is w_i. The distributions p_θs may not necessarily belong to an exponential family (e.g., the Cauchy scale family). The Fisher–Rao geometry [46,47] of the parametric family of distributions $\mathcal{F} = \{p_\theta : \theta \in \Theta\}$ (the statistical model) can be modeled as a Riemannian manifold with the Fisher metric $g(\theta) = I(\theta)$ defined by the Fisher information $I(\theta) = E_\theta[(\log p_\theta(x))^2] = -E_\theta[\nabla^2 \log p_\theta(x)]$. When $\mathcal{F}$ is an exponential family with the cumulant function $F(\theta)$, we have $I(\theta) = F''(\theta)$.

The underlying geometry of $(\mathcal{F}, g(\theta) = I(\theta))$ is Euclidean after a change in variable $\eta(\theta) = \sqrt{I(\theta)}$ since we can write the metric tensor as follows:

$$g(\theta) = \sqrt{I(\theta)} \times \underbrace{1}_{=g_{\text{Euclidean}}}, \times \sqrt{I(\theta)}.$$

Thus, the Riemannian Fisher–Rao distance is the Euclidean distance expressed in the $h(\theta)$-coordinate system with $h(\theta) = \int_{\theta_0}^{\theta} \sqrt{I(u)}\,\mathrm{d}u$, and we have the Fisher–Rao distance given by

$$\rho(p_{\theta_1}, p_{\theta_2}) = |h(\theta_1) - h(\theta_2)|.$$

When $\mathcal{F}$ is an exponential family with the cumulant function $f(\theta)$, we have $I(u) = f''(u)$.

We summarize the result on the JFR center in the following theorem:

Theorem 4 (Jeffreys–Fisher–Rao centroid in uni-order exponential families)**.** *The Jeffreys–Fisher–Rao centroid θ_S of n densities $p_{\theta_1}, \ldots, p_{\theta_n}$ of an exponential family of order one with the log-normalizer $f(\theta)$ for $\theta \in \Theta$, the natural parameter space, and weight vector $w \in \Delta_n$ is*

$$\theta_S = m_h(\bar{\theta}, \underline{\theta}), \tag{17}$$

where $m_h(\bar{\theta}, \underline{\theta}) = h^{-1}\left(\frac{h(\bar{\theta})+h(\underline{\theta})}{2}\right)$ is the quasi-arithmetic mean [35] of the dual left and right KL centroids $\bar{\theta} = \sum_{i=1}^n w_i\theta_i = \theta_R$ and $\underline{\theta} = (f')^{-1}(\sum_{i=1}^n w_i f'(\theta_i))$ with respect to the scalar monotone function $h = \int_{\theta_0}^{\theta} \sqrt{f''(u)}\,\mathrm{d}u$ for any $\theta \in \Theta$.

Proof. Since the Fisher information is $I(\theta) = f''(\theta)$, we have $h(\theta) = \int_{\theta_0}^{\theta} \sqrt{f''(u)}\mathrm{d}u$. The Riemannian center of mass [48] minimizes

$$\theta_S = \arg\min_{\theta} \sum_{i=1}^{n} w_i \rho^2(\theta_i, \theta).$$

But in the h-parameterization, the Riemannian centroid, amounts to a Euclidean center of mass/centroid in the h-Cartesian coordinate system:

$$h(\theta_S) = \sum_{i=1}^{n} w_i h(\theta_i).$$

Therefore, we have $\theta_S = h^{-1}(\sum_i w_i h(\theta_i)) =: m_h(\theta_1, \ldots, \theta_n; w_1, \ldots, w_n)$, a weighted quasi-arithmetic mean. Since the Jeffreys centroid amounts to a symmetrized Bregman centroid of the left and right Bregman centroids [28], $\underline{\theta} = m_{f'}(\theta_1, \ldots, \theta_n; w_1, \ldots, w_n)$ and $\bar{\theta} = \sum_i w_i \theta_i$. It follows that the Jeffreys–Fisher–Rao center is $\theta_{\mathrm{JFR}} = m_h(\bar{\theta}, \underline{\theta})$ after using Property 3. □

4.2. Jeffreys–Fisher–Rao Center for Categorical Distributions

Recall from Theorem 1 that the Jeffreys centroid $c = (c_1, \ldots, c_j, \ldots, c_d)$ of a set of n categorical distributions with parameters arranged in the matrix $[p_{i,j}]$ is given by

$$c_j(\lambda) = \frac{a_j}{W_0\left(\frac{a_j}{g_j} e^{1+\lambda}\right)}, \quad \forall j \in \{1, \ldots, d\},$$

where $a_j = \sum_{i=1}^{n} w_i p_{i,j}$ and $g_j = \frac{\prod_{i=1}^{n} p_{i,j}^{w_i}}{\sum_{j=1}^{d} \prod_{i=1}^{n} p_{i,j}^{w_i}}$ are the components of the weighted arithmetic and normalized geometric means, respectively, and W_0 is the principal branch of the Lambert W function [33]. The optimal $\lambda \le 0$ is unique and satisfies $\lambda = -D_{\mathrm{KL}}(c_j(\lambda) : g)$.

Let $c(\lambda) = (c_1(\lambda), \ldots, c_d(\lambda))$. Let $L_J(p)$ denote the Jeffreys loss function to minimize to find the optimal Jeffreys centroid:

$$L_J(p) = \sum_{i=1}^{n} w_i D_J(p_i, p) \tag{18}$$

We say that p is a $(1+\epsilon)$ approximation of the exact Jeffreys centroid c when we have

$$L_J(c) \le L_J(p) \le (1+\epsilon) L_J(c).$$

It was shown in [31] that $\tilde{c} = c(0)$, called the unnormalized Jeffreys center, yields a $s(\lambda) - 1$ approximation on c where $s(\lambda) = \sum_j c_j(\lambda) \le 1$.

Since the Fisher–Rao geodesic midpoints on the categorical Fisher–Rao manifold are known in closed form [49], we give the mathematical expression of the JFR center as follows:

Theorem 5 (JFR centroid of categorical distributions)**.** *Let $\mathcal{P}_w = \{p_1, \ldots, p_n\}$ be a set of n probability mass functions weighted by $w \in \Delta_n$ with $p_i = (p_{i,1}, \ldots, p_{i,d}) \in \Delta_d$ for $i \in \{1, \ldots, n\}$ and $w \in \Delta_n$. Then, the JFR barycenter c minimizing is unique and given by the following formula:*

$$c_j = \frac{(\sqrt{a_j} + \sqrt{g_j})^2}{2\left(1 + \sum_{l=1}^{d} \sqrt{a_j}\sqrt{g_j}\right)}, \forall j \in \{1, \ldots, d\}, \tag{19}$$

where $a = (a_1, \ldots, a_d) = \sum_{i=1}^{n} w_i p_i$ is the weighted arithmetic mean and $g = (g_1, \ldots, g_d)$ is the normalized weighted geometric mean with components $g_j = \frac{\prod_{i=1}^{n} p_{i,j}^{w_i}}{\sum_{j=1}^{d} \prod_{i=1}^{n} p_{i,j}^{w_i}}$ for $i \in \{1, \ldots, d\}$.

Notice that the JFR center differs from the Jeffreys centroid, which requires the use of the Lambert W function [33]. However, we noticed that for practical applications, the JFR centroid approximates the Jeffreys centroid well and is much faster to compute (see the experiments in Section 5).

4.3. Jeffreys–Fisher–Rao Center for Multivariate Normal Distributions

Let $\mathcal{P} = \{p_{\mu_1,\Sigma_1}, \ldots, p_{\mu_n,\Sigma_n}\}$ be a set of n probability density functions (PDFs) of d-variate normal distributions weighted by $w \in \Delta_n$, where the PDF of a multivariate normal distribution of mean μ and the covariance matrix Σ is given by

$$p_{\mu,\Sigma} = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Let $\lambda_i = (\mu_i, \Sigma_i)$ be the ordinary parameterization of normal distributions p_{μ_i,Σ_i}. The family $\mathcal{F} = \{p_{\mu,\Sigma}(x) : \mu \in \mathbb{R}^d, \Sigma \in \mathrm{Sym}^{++}(\mathbb{R}, d)\}$ of multivariate normal distributions forms an exponential family with the dual natural θ- and moment η-parameterizations [7] given by

$$\begin{aligned} \theta(\lambda) &= (\theta_v, \theta_M) = \left(\Sigma^{-1}\mu, \frac{1}{2}\Sigma^{-1}\right), \\ \eta(\lambda) &= \left(\mu, \mu\mu^\top + \Sigma\right), \end{aligned}$$

when choosing the sufficient statistic $t(x) = (x, xx^\top)$. The Jeffreys divergence between two d-variate normal distributions $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ is given by the formula

$$D_J(p_{\mu_1,\Sigma_1}, p_{\mu_2,\Sigma_2}) = (\mu_2 - \mu_1)\top(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_2 - \mu_1) + \mathrm{tr}\left(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1\right) - 2d.$$

The left and right Kullback–Leibler barycenters amount to the corresponding right and left Bregman barycenters [28] induced by the cumulant function

$$F(\theta) = F(\theta_v, \theta_M) = \frac{1}{2}\left(d \log \pi - \log\det(\theta_M) + \frac{1}{2}\theta_v^\top \theta_M^{-1}\theta_v\right),$$

and the gradient of $F(\theta)$ defines the dual moment parameter with

$$\eta(\theta) = \nabla F(\theta) = \left(\frac{1}{2}\theta_M^{-1}\theta_v, \frac{1}{2}\theta_M^{-1} - \frac{1}{4}(\theta_M^{-1}\theta_v)(\theta_M^{-1}\theta_v)^\top\right).$$

The reciprocal gradient is given by

$$\theta(\eta) = \theta(\eta_v, \eta_M) = (\nabla F)^{-1}(\eta) = \left(\theta_v = -(\eta_M + \eta_v\eta_v^\top)^{-1}\eta_v, \theta_M = -\frac{1}{2}(\eta_M + \eta_v\eta_v^\top)^{-1}\right).$$

The Gauss–Bregman center is a $(A, m_{\nabla F})$-inductive center, which can be approximated by carrying a prescribed number T of iterations of the Gauss–Bregman double sequence.

Although the Rao distance between two d-variate normal distributions is not available in closed form when $d > 1$ [50,51], the Jeffreys–Fisher–Rao center can be computed in closed form. Indeed, the sided Kullback–Leibler centroids of multivariate normal distributions amount to reverse-sided Bregman centroids [28], and the Fisher–Rao geodesic midpoint between two multivariate normal distributions was recently reported in [32]. Appendix C concisely describes the method of Kobayashi [32], which allows one to obtain the Fisher–Rao midpoints of multivariate normal distributions. An implementation of that algorithm is available in the Python software library `pyBregMan` [52].

Thus, the Jeffreys–Fisher–Rao center is available in closed form:

Theorem 6 (JFR center of MVNs)**.** *The Jeffreys–Fisher–Rao center of a finite set of weighted multivariate normal distributions is available in closed form.*

Note that the Fisher–Rao distance between normal distributions is invariant under the action of the positive affine group [50], as are the Jeffreys centroid, the JFR center, and the GB center. Figure 6 shows several examples of the JFR and GB centers of two univariate normal distributions. We can observe that those centers are close to each other although they are distinct when the normal distributions do not share the same means and covariance matrices.

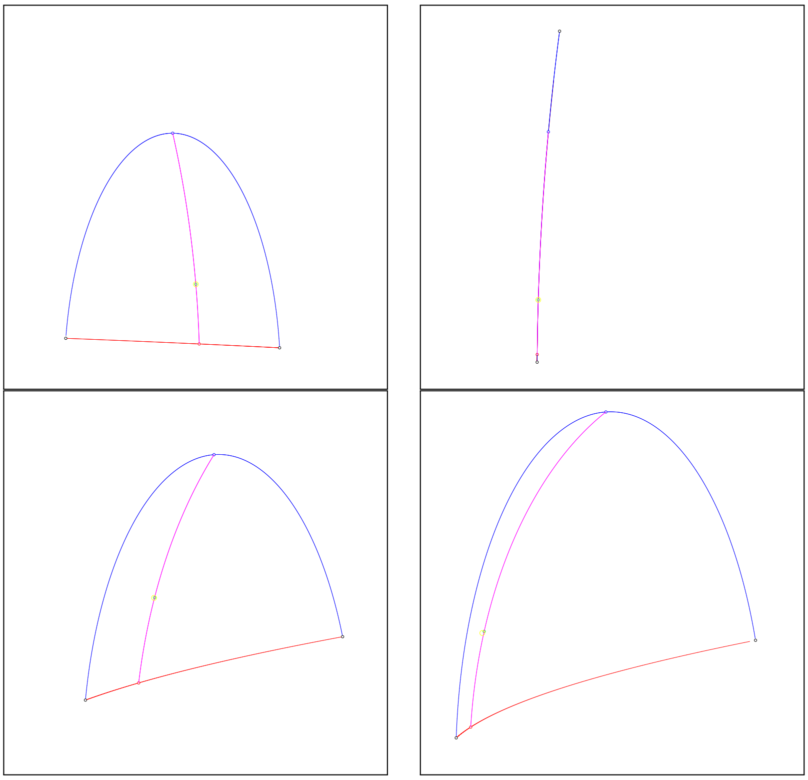

Figure 6. Visualization of the Jeffreys–Fisher–Rao center and Gauss–Bregman center of two univariate normal distributions (black circle). The exponential geodesic and mixture geodesics are shown in red and blue, respectively, with their corresponding midpoints. The Jeffreys–Fisher–Rao is the Fisher–Rao midpoint (green) lying on the Fisher–Rao geodesics (purple). The inductive Gauss–Bregman center is displayed in yellow with double size in order to ease its comparison with the Jeffreys–Fisher–Rao center.

Figure 7 shows the various centroids/centers between two bivariate normal distributions displayed as ellipsoids centered as their means. Observe that the inductive Gauss–Bregman center is visually closer than the Jeffreys–Fisher–Rao center to the Jeffreys centroid.

Figure 8 displays the various centroids and centers for pairs of bivariate normal distributions centered at the same mean. Figure 9 shows the centroids and centers for pairs of bivariate normal distributions with the same covariance matrix.

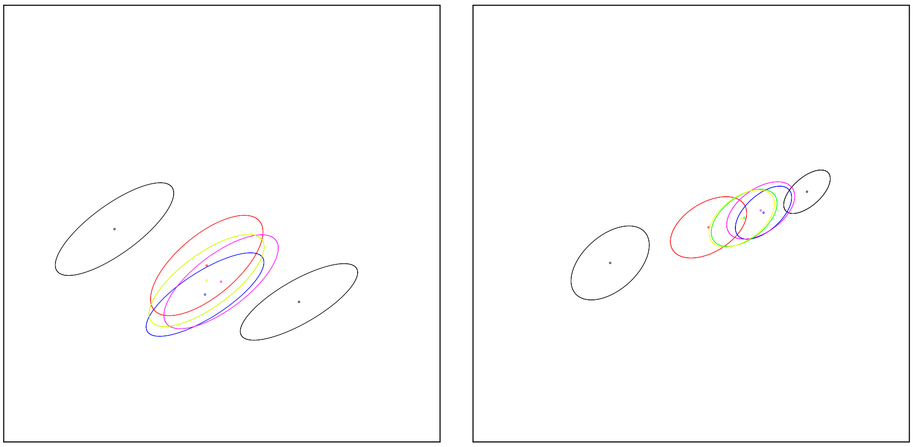

Figure 7. Centroids and centers between a pair of bivariate normal distributions (black). Each normal distribution $N(\mu, \Sigma)$ (parameterized by a 5D parameter θ) is displayed as a 2D ellipsoid $\mathcal{E}(\mu, \Sigma) = \{(x-\mu)^\top \Sigma^{-1}(x-\mu) = l\}$ for a prescribed level $l > 0$ in the sample space $\mathbb{R}^2$. Blue, red, purple, yellow, and green ellipsoids correspond to m-geodesic midpoint, e-geodesic midpoint, Jeffreys–Fisher–Rao midpoint, Gauss–Bregman inductive mean, and numerical Jeffreys centroid (symmetrized Bregman centroid), respectively.

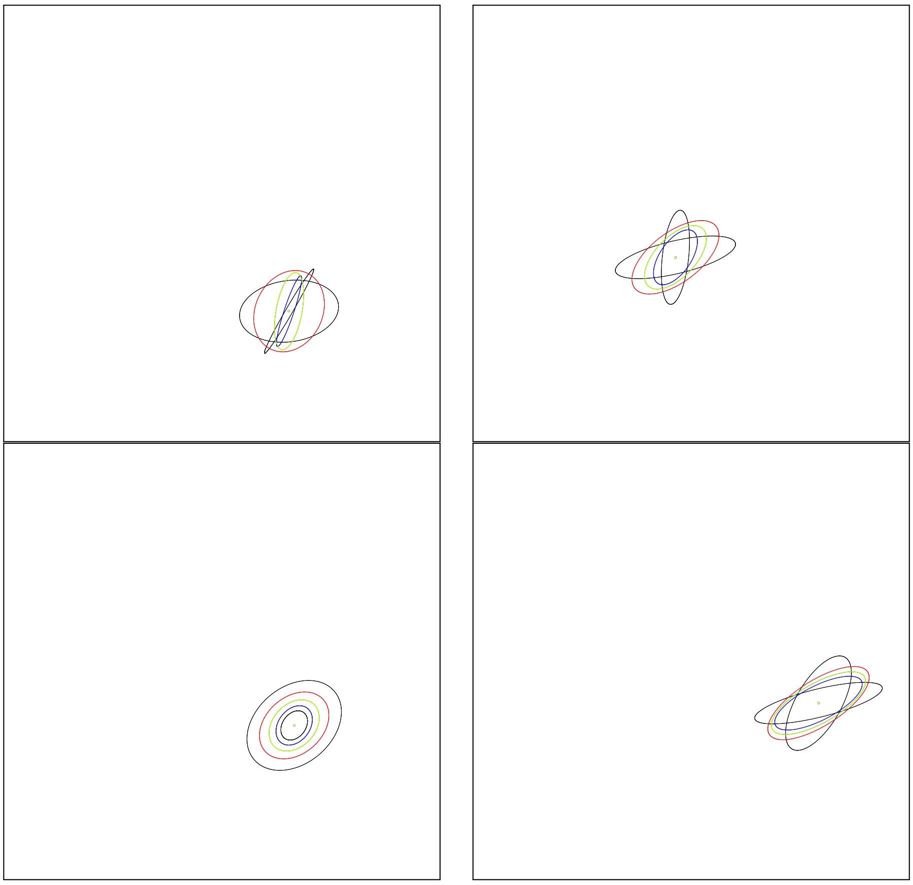

Figure 8. Centroids and centers between a pair of bivariate centered normal distributions (black). Each normal distribution $N(\mu, \Sigma)$ with a prescribed μ (parameterized by a 3D parameter θ) is displayed as a 2D ellipsoid. The red and blue ellipsoids correspond to the e-geodesic and m-geodesic midpoints, respectively. The green ellipsoid is the exact Jeffreys centroid which coincide perfectly with the inductive Gauss–Bregman center (yellow) and Jeffreys–Fisher–Rao center (purple). Thus these three green, yellow, and purple matching ellipsoids are rendered superposed in an overall shade of brown.

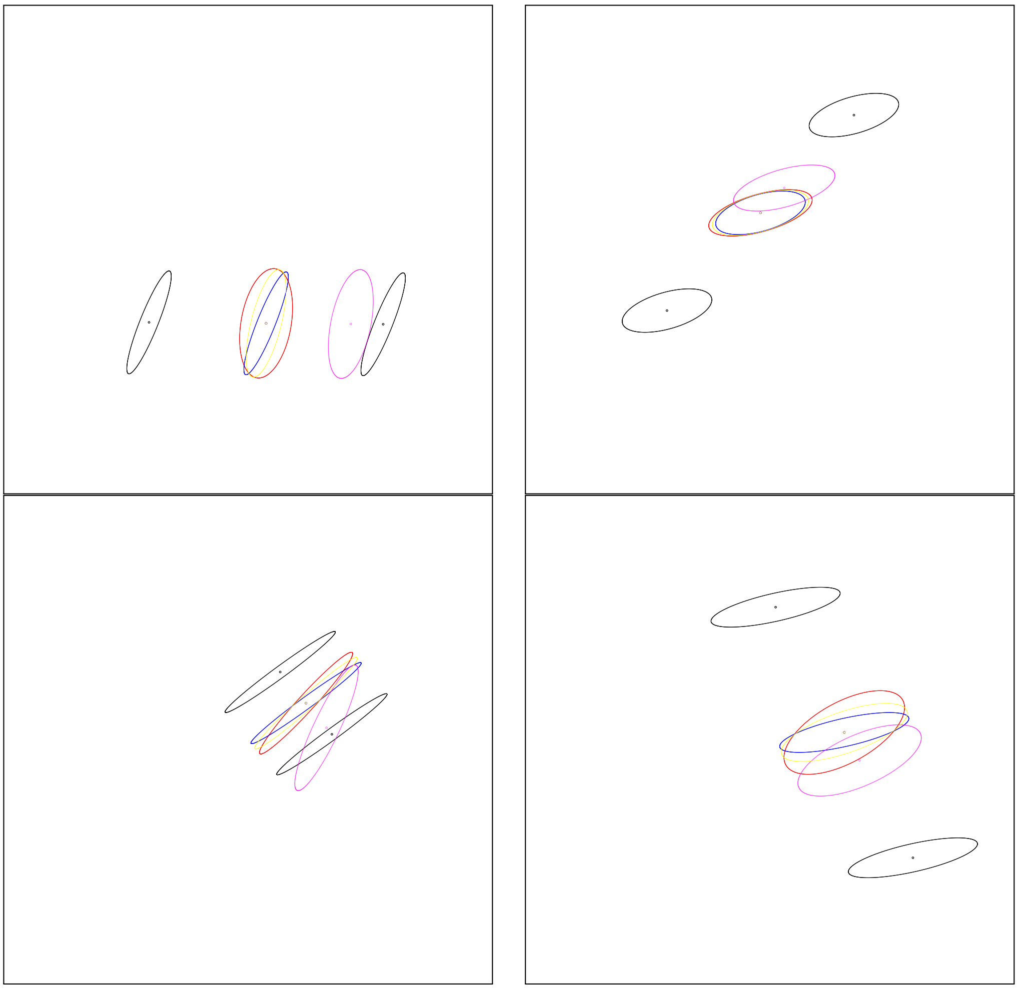

Figure 9. Centroids and centers between a pair of bivariate-centered normal distributions (black). Each normal distribution $N(\mu, \Sigma)$ with a prescribed covariance matrix Σ (parameterized by a 2D parameter θ) is displayed as a 2D ellipsoid. The red and blue ellipsoids correspond to the e-geodesic and m-geodesic midpoints, respectively. The inductive Gauss–Bregman (yellow) and Jeffreys–Fisher–Rao center (purple) do not coincide.

Remark 1. *In general, an exponential family may be characterized equivalently by two convex functions: (1) its log-normalizer $F(\theta)$ or (2) its partition function $Z(\theta) = \exp(F(\theta))$, which is log-convex and hence also convex [53]. It has been shown that the Bregman divergence B_Z for $Z = \sqrt{\det(\theta)}$ (convex) corresponds to the reverse extended Kullback–Leibler divergence between unnormalized PDFs of normal distributions:*

$$B_Z(\theta_1 : \theta_2) = D^+_{\mathrm{KL}}(\tilde{p}_{\lambda(\theta_2)} : \tilde{p}_{\lambda(\theta_1)}),$$

where $\tilde{p}_{\mu,\Sigma} = \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$ and the extended KLD between two positive measures is given by

$$D^+_{\mathrm{KL}}(m_1 : m_2) = \int \left(m_1(x) \log \frac{m_1(x)}{m_2(x)} + m_2(x) - m_1(x) \right) \mathrm{d}\mu(x).$$

Remark 2. *We may further define yet another center for multivariate normal distributions by considering the Fisher–Rao isometric embedding of the Fisher–Rao d-variate normal manifold $\mathcal{M} = \{p_{\mu,\Sigma}\}$ into the Fisher–Rao $(d+1)$-variate centered manifold $\mathcal{N}_0^+ = \{q_P(y) = p_{0,P}(y) : P \in \mathrm{Sym}^{++}(\mathbb{R}, d+1)\}$ using Calvo and Oller mapping [50]:*

$$f(\mu,\Sigma) := \begin{bmatrix} \Sigma + \mu\mu^\top & \mu \\ \mu^\top & 1 \end{bmatrix}.$$

Let $\bar{\mathcal{M}} = \{f(p) : p \in \mathcal{M}\}$ denote the embedded submanifold of codimension one in $\mathcal{N}_0^+$. The Calvo–Oller center is then defined by taking the Fisher–Rao midpoints q_{CO} of q_{P_1} and q_{P_2}, projecting q_{CO} onto $\bar{\mathcal{M}}$ as q'_{CO} and converting q'_{CO} into $p_{\mathrm{CO}} \in \mathcal{M}$ using the inverse mapping f^{-1} [51].

The Fisher orthogonal projection of a $(d+1)\times(d+1)$ matrix $P \in \mathcal{N}_0^+$ onto the submanifold $\bar{\mathcal{M}}$ is performed as follows: Let $\beta = P_{d+1,d+1}$ and write $P = \begin{bmatrix} \Sigma + \beta\mu\mu^\top & \beta\mu \\ \beta\mu^\top & \beta \end{bmatrix}$. Then, the orthogonal projection at $P \in \mathcal{P}$ onto $\bar{\mathcal{M}}$ is $\begin{bmatrix} \Sigma + \mu\mu^\top & \mu^\top \\ \mu & 1 \end{bmatrix}$. See [51] for details of the Calvo and Oller embedding/projection method.

5. Experiments

We run all experiments on a Dell Inspiron 5502 Core i7-116567@2.8Ghz using compiled Java programs. For each experiment, we consider a set of $n = 2$ uniformly randomized histograms with d bins (i.e., points in Δ_d) and calculate the numerical Jeffreys centroid, which requires the time-consuming Lambert W function, the GB center, and the JFR center. For each prescribed value of d, we run 10000 experiments to collect various statistics like the average and maximum approximations and running times. The approximations of the JFR and GB methods are calculated either as the approximation of the Jeffreys information (Equation (18)) or as the approximation of the centers with respect to the numerical Jeffreys centroids measured using the total variation distance. Table 1 is a verbatim export of our experimental results when we range the dimension of histograms for $d = 2$ to $d = 256$ by doubling the dimension at each round. The inductive GB center is stopped when the total variation $\frac{1}{2}\|a_t - g_t\|_1 \leq 10^{-8}$.

We observe that the JFR center is faster to compute than the GB center but the GB center is of higher quality (i.e., a better approximation with a lower ϵ) than the JFR center to approximate the numerical Jeffreys centroid.

Another test consists of choosing $d = 3$ and the following two 3D normalized histograms: $(\frac{1}{3},\frac{1}{3},\frac{1}{3})$ and $(1-\alpha, \alpha/2, \alpha/2)$ for $\alpha \in \{10^{-1}, 10^{-2}, \ldots, 10^{-7}, 10^{-8}\}$. Table 2 reports the experiments. The objective is to find a setting where both the JFR and GB centers are distinguished from the Jeffreys centroid. We see that as we decrease α, the approximation factor ϵ gets worse for both the JFR center and the GB center. The JFR center is often faster to compute than the GB inductive center, but the approximation of the GB center is better than the JFR approximation.

Finally, we implemented the Gauss–Bregman and Jeffreys–Fisher–Rao centers and Jeffreys centroid using multi-precision arithmetic. We report the following experiments using 200-digit precision arithmetic for the following input of two normalized histograms: $p = (0.1, 0.9)$ and $q = (0.8, 0.2)$. We report the various first 17-digit mantissas obtained with the corresponding Jeffreys information:

- Jeffreys center: $(0.42490383904214813, 0.575096160957851866)$
  Jeffreys information: 1.2490723231955352.
- Gauss–Bregman center: $(\mathbf{0.424903839042}76856, \mathbf{0.575096160957}231439)$
  Jeffreys information: 1.2490723231955353.
- Jeffreys–Fisher–Rao center: $(\mathbf{0.42490390}202906282, \mathbf{0.575096}097970937175)$
  Jeffreys information: 1.2490723232068266.

The total variation distance between the Jeffreys centroid and the Gauss–Bregman center is 6.2042711482840874223500006863 72 10^{-13}.

The total variation distance between the Jeffreys centroid and the Jeffreys–Fisher–Rao center is 6.29869146904791198403936376261 1 10^{-8}.

Table 1. Experiments for JFR and GB centers approximating the numerical Jeffreys centroid.

dim.			**Jeffreys–Fisher–Rao Center**						**Gauss–Bregman Center**			
	Avg Info ϵ	**Max Info ϵ**	**Avg TV**	**Max TV**	**Avg Time**	**× speed**	**Avg Info ϵ**	**Max Info ϵ**	**Avg TV**	**Max TV**	**Avg Time**	**× Speed**
d = 2	5.662×10^{-6}	6.386×10^{-3}	8.735×10^{-5}	5.005×10^{-2}	1.614×10^{-7}	82.541	1.507×10^{-4}	9.745×10^{-2}	6.304×10^{-4}	5.005×10^{-2}	5.072×10^{-7}	26.258
d = 4	1.283×10^{-5}	5.294×10^{-3}	1.690×10^{-4}	3.969×10^{-2}	1.418×10^{-7}	182.309	4.696×10^{-4}	7.695×10^{-2}	1.431×10^{-3}	3.969×10^{-2}	1.623×10^{-7}	159.304
d = 8	2.766×10^{-5}	6.970×10^{-3}	2.210×10^{-4}	3.470×10^{-2}	1.772×10^{-7}	292.125	1.011×10^{-3}	9.677×10^{-2}	2.033×10^{-3}	3.470×10^{-2}	1.955×10^{-7}	264.680
d = 16	3.531×10^{-5}	8.544×10^{-3}	2.325×10^{-4}	2.450×10^{-2}	6.318×10^{-7}	224.370	1.388×10^{-3}	9.231×10^{-2}	2.275×10^{-3}	2.450×10^{-2}	7.208×10^{-7}	196.660
d = 32	4.123×10^{-5}	5.242×10^{-3}	2.457×10^{-4}	1.230×10^{-2}	4.811×10^{-7}	462.754	1.674×10^{-3}	5.398×10^{-2}	2.449×10^{-3}	1.230×10^{-2}	5.457×10^{-7}	408.007
d = 64	4.747×10^{-5}	3.437×10^{-3}	2.486×10^{-4}	9.756×10^{-3}	9.789×10^{-7}	578.354	1.863×10^{-3}	3.685×10^{-2}	2.498×10^{-3}	9.756×10^{-3}	1.160×10^{-6}	488.246
d = 128	5.020×10^{-5}	2.540×10^{-3}	2.491×10^{-4}	6.580×10^{-3}	5.874×10^{-6}	477.412	1.937×10^{-3}	2.374×10^{-2}	2.522×10^{-3}	6.580×10^{-3}	6.605×10^{-6}	424.609
d = 256	4.735×10^{-5}	1.410×10^{-3}	2.476×10^{-4}	4.855×10^{-3}	9.349×10^{-6}	528.452	1.914×10^{-3}	1.521×10^{-2}	2.529×10^{-3}	4.855×10^{-3}	1.110×10^{-5}	445.304

Table 2. Experiments for JFR and GB centers approximating the numerical Jeffreys centroid for the following setting of two normalized histograms of 3 bins: $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $(1-\alpha, \alpha/2, \alpha/2)$.

α	**Info. ϵ**	**TV ϵ**	**Avg Time**	**× Speed**	**Info. ϵ**	**TV ϵ**	**Avg Time**	**× Speed**
1.000×10^{-1}	6.882×10^{-9}	2.495×10^{-5}	1.767×10^{-7}	125.960	1.338×10^{-6}	3.480×10^{-4}	2.334×10^{-7}	95.356
1.000×10^{-2}	2.607×10^{-5}	1.722×10^{-3}	1.371×10^{-7}	167.932	1.061×10^{-3}	1.108×10^{-2}	1.565×10^{-7}	147.104
1.000×10^{-3}	6.262×10^{-4}	7.530×10^{-3}	1.033×10^{-7}	218.450	1.272×10^{-2}	3.534×10^{-2}	1.208×10^{-7}	186.698
1.000×10^{-4}	3.632×10^{-3}	1.570×10^{-2}	1.171×10^{-7}	193.345	4.580×10^{-2}	6.065×10^{-2}	1.367×10^{-7}	165.571
1.000×10^{-5}	1.121×10^{-2}	2.419×10^{-2}	1.546×10^{-7}	150.807	7.322×10^{-3}	1.929×10^{-2}	2.834×10^{-7}	82.261
1.000×10^{-6}	2.457×10^{-2}	3.204×10^{-2}	1.619×10^{-7}	141.896	1.655×10^{-2}	2.579×10^{-2}	2.512×10^{-7}	91.467
1.000×10^{-7}	4.375×10^{-2}	3.897×10^{-2}	1.357×10^{-7}	170.065	3.065×10^{-2}	3.183×10^{-2}	2.131×10^{-7}	108.314
1.000×10^{-8}	6.806×10^{-2}	4.492×10^{-2}	1.315×10^{-7}	173.698	4.948×10^{-2}	3.725×10^{-2}	2.017×10^{-7}	113.292
1.000×10^{-9}	9.651×10^{-2}	4.999×10^{-2}	1.125×10^{-7}	208.627	7.240×10^{-2}	4.199×10^{-2}	1.590×10^{-7}	147.610
1.000×10^{-10}	1.281×10^{-1}	5.428×10^{-2}	8.366×10^{-8}	242.967	9.862×10^{-2}	4.610×10^{-2}	1.111×10^{-7}	183.000
1.000×10^{-11}	1.621×10^{-1}	5.792×10^{-2}	1.066×10^{-7}	215.817	1.274×10^{-1}	4.963×10^{-2}	1.325×10^{-7}	173.614
1.000×10^{-12}	1.979×10^{-1}	6.100×10^{-2}	1.028×10^{-7}	229.484	1.580×10^{-1}	5.266×10^{-2}	1.329×10^{-7}	177.581
1.000×10^{-13}	2.348×10^{-1}	6.363×10^{-2}	9.541×10^{-8}	244.587	1.901×10^{-1}	5.526×10^{-2}	1.255×10^{-7}	185.940
1.000×10^{-14}	2.727×10^{-1}	6.589×10^{-2}	1.062×10^{-7}	219.787	2.231×10^{-1}	5.750×10^{-2}	1.361×10^{-7}	171.456
1.000×10^{-15}	3.112×10^{-1}	6.784×10^{-2}	9.043×10^{-8}	248.688	2.570×10^{-1}	5.943×10^{-2}	1.322×10^{-7}	170.122
1.000×10^{-16}	3.483×10^{-1}	6.946×10^{-2}	8.857×10^{-8}	267.219	2.897×10^{-1}	6.105×10^{-2}	1.438×10^{-7}	164.535

The total variation distance between the Gauss–Bregman center and the Jeffreys–Fisher–Rao center is 6.298629426336429143165140262604 10^{-8}.

Although all those points are close to each other, they are all distinct points (note that using the limited precision of the IEEE 754 floating point standard may yield a misleading interpretation of experiments).

6. Conclusions and Discussion

In this work, we considered the Jeffreys centroid of a finite weighted set of densities of a given exponential family $\mathcal{E} = \{p_\theta(x) : \theta \in \Theta\}$. This Jeffreys centroid amounts to a symmetrized Bregman centroid on the corresponding weighted set of natural parameters of the densities [28]. In general, the Jeffreys centroids do not admit closed-form formulas [28,31] except for sets of same-mean normal distributions [29] (see Appendix B).

In this paper, we interpreted the closed-form formula for same-mean multivariate normal distributions in two different ways:

- First, as the Fisher–Rao geodesic midpoint of the sided Kullback–Leibler centroids. This interpretation lets us relax the midpoint definition to arbitrary exponential families to define the Jeffreys–Fisher–Rao center (the JFR center of Definition 2);
- Second, as an inductive $(A, m_{\nabla F})$ center using a multivariate Gauss-type double sequence, which converges to the Gauss–Bregman center (the GB center of Definition 3). The latter definition yields an extension of Nakamura's arithmetic–harmonic (A, H) mean [41] to an arbitrary $(A, m_{\nabla F})$ mean for which we proved convergence under a separability condition in Theorem 3. Convergence proof remains to be performed in the general case, although we noticed in practice convergence when $\nabla F(\theta)$ is the moment parameter of categorical or multivariate normal distributions.

In general, the Jeffreys, JFR, and GB centers differ from each other (e.g., the case of categorical distributions). But for sets of same-mean normal distributions, they remarkably coincide altogether: namely, this was the point of departure of this research. We reported generic or closed-form formulas for the JFR centers of (a) uni-order parametric families in Section 4.1 (Theorem 4), (b) categorical families in Section 4.2 (Theorem 5), and (c) multivariate normal families in Section 4.3 (Theorem 6). Table 3 summarizes the new results obtained in this paper and states references of prior work. Notice that in practice, we approximate the Gauss–Bregman center by prescribing a number of iterations $T \in \mathbb{N}$ for the Gauss–Bregman double sequence to obtain $m_{\mathrm{GB}}^{(T)}$. Prescribing the number of GB center iterations T allows us to tune the time complexity of computing $m_{\mathrm{GB}}^{(T)}$ while adjusting the quality of the approximation of the Jeffreys centroid.

Table 3. Summary of the results: $\triangle$ indicates a generic formula, $\surd$ a closed-form formula, and $\times$ no-known formula.

Family	Jeffreys	Jeffreys–Fisher–Rao	Gauss–Bregman
Exponential family	Equation (2)	Definition 2	Definition 3
One-dimensional exponential family	$\times$ Theorem 4	$\triangle$	$\times$ [43]
Categorical family	$\triangle$ [31]	$\surd$ Theorem 5	$\times$ Theorem 3
Normal family	$\times$ [28]	$\surd$ Theorem 6	$\times$ Theorem 3
Centered normal family	$\surd$ [29]	$\surd$ [29]	$\surd$ [41]

In applications requiring the Jeffreys centroid, we thus propose to either use the fast Jeffreys–Fisher–Rao center when a closed-form formula is available for the family of distributions at hand or use the Gauss–Bregman center approximation with a prescribed number of iterations as a drop-in replacement of the numerical Jeffreys centroids while keeping the Jeffreys divergence (the centers we defined are not centroids as we do not exhibit distances from which they are population minimizers).

More generally, let us rephrase the results in a purely geometric setting using the framework of information geometry [14]: let $P_1, \ldots, P_n$ be a set of n points weighted by a vector $w \in \Delta_n$ on an m-dimensional dually flat space (M, g, ∇, ∇^*) with the ∇-affine coordinate system $\theta(\cdot)$ and dual ∇^*-affine coordinate system $\eta(\cdot)$, where ∇ and ∇^* are two torsion-free dual affine connections. The Riemannian metric g is a Hessian metric [54], which may be expressed in the θ-coordinate system as $g(\theta) = \nabla^2 F(\theta)$ or in the dual coordinate system as $g(\eta) = \nabla^2 F^*(\eta)$, where $F(\theta)$ and $F^*(\eta)$ are dual convex potential functions related by the Legendre–Fenchel transform [14,54]. Let $\eta_i = \nabla F(\theta_i)$ and $\theta_i = \nabla F^*(\eta_i)$ be the coordinates of point P_i in the η- and θ-coordinate systems, respectively. An arbitrary point P can be either referenced in the θ-coordinate system ($P = P_\theta$) or in the η-coordinate system ($P = P_\eta$). Then, the Jeffreys–Fisher–Rao center is defined as the midpoint with respect to the Levi-Civita connection $\bar{\nabla} = \frac{\nabla + \nabla^*}{2} = \nabla^g$ of g:

$$C_{\mathrm{JFR}} := \gamma_{\bar{\nabla}}(C_{\bar{\theta}}, C_{\underline{\theta}}, \frac{1}{2}) =: C_{\bar{\theta}} \# C_{\underline{\theta}}. \tag{20}$$

The point $C_{\bar{\theta}}$ is the centroid with respect to the canonical flat divergence $\mathcal{D}(P : Q) = F(\theta(P)) + F^*(\eta(Q)) - \sum_{i=1}^m \theta_i(P)\eta_i(Q)$, and the point $C_{\underline{\theta}}$ is the centroid with respect to the dual canonical flat divergence $\mathcal{D}^*(P : Q) := \mathcal{D}(Q : P)$. The canonical divergence is expressed using the mixed coordinates θ/η but can also be expressed using the θ-coordinates as an equivalent Bregman divergence $\mathcal{D}(P : Q) = B_F(\theta(P) : \theta(Q))$ or as a reverse dual Bregman divergence $\mathcal{D}(P : Q) = B_{F^*}(\eta(Q) : \eta(P))$. This JFR center C_{JFR} approximates the symmetrized centroid with respect to the canonical symmetrized divergence $\mathcal{S}(P, Q) = \mathcal{D}(P : Q) + \mathcal{D}(Q : P)$ (i.e., Jeffreys divergence when written using the θ-coordinate system). This symmetrized divergence $\mathcal{S}(P, Q)$ can be interpreted as the energy of the Riemannian length element $\mathrm{d}s$ along the primal geodesic $\gamma(t)$ and dual geodesic $\gamma^*(t)$ (with $\gamma(0) = \gamma^*(0) = P$ and $\gamma(1) = \gamma^*(1) = Q$), see [14]: $\mathcal{S}(P, Q) = \int_0^1 \mathrm{d}s^2(\gamma(t))\mathrm{d}t = \int_0^1 \mathrm{d}s^2(\gamma^*(t))\mathrm{d}t$. The Riemannian distance $\rho(P, Q)$ corresponds to the Riemannian length element along the Riemannian geodesic $\bar{\gamma}(t)$ induced by the Levi-Civita connection $\bar{\nabla} = \frac{\nabla + \nabla^*}{2}$: $\rho(P, Q) = \int_0^1 \mathrm{d}s(\bar{\gamma}(t))\mathrm{d}t$.

The inductive Gauss–Bregman center C_{GB} is obtained as a limit sequence of taking iteratively the ∇ midpoints and ∇^* midpoints with respect to the ∇ and ∇^* connections. Those midpoints correspond to the right and left centroids C_{t+1} and C_{t+1}^* with respect to $\mathcal{D}(\cdot : \cdot)$:

$$\begin{aligned} C_{t+1} &= \gamma_{\nabla}\left(C_t, C_t^*, \frac{1}{2}\right), \\ C_{t+1}^* &= \gamma_{\nabla^*}\left(C_t, C_t^*, \frac{1}{2}\right), \end{aligned}$$

initialized with $\theta(C_0) = \sum_{i=1}^n w_i \theta(P_i)$ and $\eta(C_0^*) = \sum_{i=1}^n w_i \eta(P_i)$. We have $C_0 = \arg\min_{C \in M} \sum_i w_i \mathcal{D}(P_i : C)$ and $C_0^* = \arg\min_{C \in M} \sum_i w_i \mathcal{D}(P_i : C)$. Figure 10 geometrically illustrates the double sequence of iteratively taking dual geodesic midpoints to converge toward the Gauss–Bregman center C_{GB}. Thus, the GB double sequence can be interpreted as a geometric optimization technique. Figure 11 illustrates the JFR and GB centers on a dually flat space. Notice that C_{JFR} has coordinates $\mathrm{JFR}_F(\mathcal{P}_\theta; w)$ in the θ-chart and coordinates $\mathrm{JFR}_{F^*}(\mathcal{P}_\eta; w)$ in the η-chart. Similarly, C_{GB} has coordinates $\mathrm{GB}_F(\bar{\theta}, \underline{\theta})$ in the θ-chart and coordinates $\mathrm{GB}_{F^*}(\bar{\eta}, \underline{\eta})$ in the η-chart.

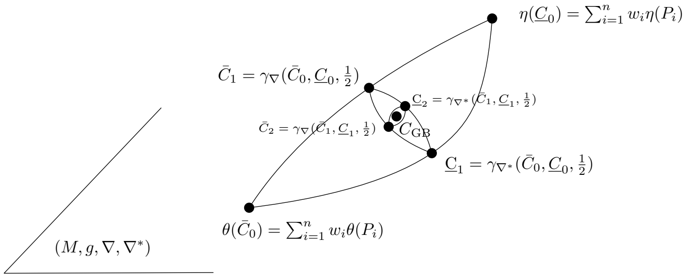
$\eta(\underline{C}_0) = \sum_{i=1}^n w_i \eta(P_i)$
$\bar{C}_1 = \gamma_\nabla(\bar{C}_0, \underline{C}_0, \frac{1}{2})$
$\underline{C}_2 = \gamma_{\nabla^*}(\bar{C}_1, \underline{C}_1, \frac{1}{2})$
$\bar{C}_2 = \gamma_\nabla(\bar{C}_1, \underline{C}_1, \frac{1}{2})$
C_{GB}
$\underline{C}_1 = \gamma_{\nabla^*}(\bar{C}_0, \underline{C}_0, \frac{1}{2})$
$\theta(\bar{C}_0) = \sum_{i=1}^n w_i \theta(P_i)$
(M, g, ∇, ∇^*)

Figure 10. Illustration on a dually flat space of the double sequence inducing a Gauss–Bregman center in the limit.

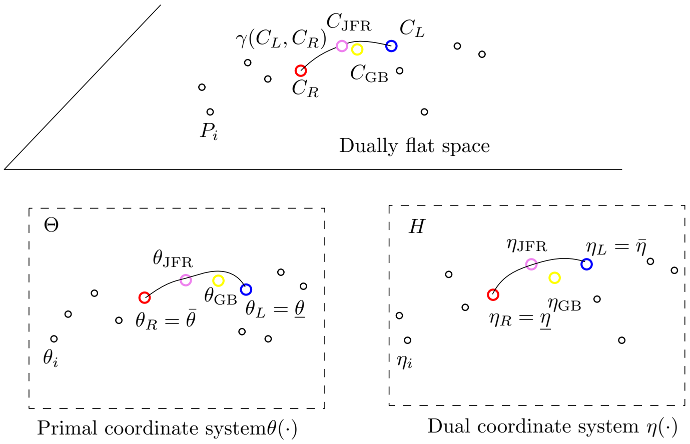

Figure 11. Illustration of the Jeffreys–Fisher–Rao and Gauss–Bregman centers a dually flat space. γ denotes the Riemannian geodesic.

As a final remark, let us emphasize that choosing a proper mean or center depends on the application at hand [55,56]. For example, in Bayesian hypothesis testing, the Chernoff mean [57] is used to upper bound Bayes' error and has been widely used in information fusion [18] for its empirical robustness [58] in practice. Jeffreys centroid has been successfully used in information retrieval tasks [6].

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author Frank Nielsen is employed by the company Sony Computer Science Laboratories Inc. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The authors declare no conflicts of interest.

Appendix A. Numerical Jeffreys Centroids for Categorical Distributions

Algorithm A1 implements the method described in [31] for numerically finely approximating the Jeffreys centroid of a weighted set of categorical distributions.

Algorithm A1: Numerical approximation of the SKL/Jeffreys centroid for categorical distributions.

```
Input: A set of weighted categorical distributions: $\mathcal{P}^w = \{p_1, \ldots, p_n\}$ with
        $w \in \Delta_n$ and $p_i \in \Delta_d$. Let $p_{i,j}$ denote the $j$-th component of $p_i$.
Input: A precision parameter $\epsilon > 0$
Output: A numerical approximation of the SKL centroid/Jeffreys centroid $c$
/* Arithmetic weighted mean (normalized)                                        */
$a_j = \sum_{i=1}^n w_i p_{i,j}$ for $i \in \{1, \ldots, d\}$ for $j \in \{1, \ldots, d\}$ ;
/* Normalized geometric weighted mean                                            */
$g_j = \frac{\prod_{i=1}^n p_{i,j}^{w_i}}{\sum_{j=1}^d \prod_{i=1}^n p_{i,j}^{w_i}}$ for $j \in \{1, \ldots, d\}$ ;
/* Initialize range where to find the optimal $\lambda^*$                         */
$\lambda_M = 0$; $\lambda_m = \max_{i \in \{1,\ldots,n\}} \{a_i + \log g_i\} - 1$ ;
/* Bisection search                                                              */
while $|\lambda_M - \lambda_m| > \epsilon$ do
    $\lambda = \frac{\lambda_m + \lambda_M}{2}$ ;
    /* $W_0$ is the principal branch of Lambert $W$ function                      */
    $c_j(\lambda) = \frac{a_j}{W_0\left(\frac{a_j}{g_j} e^{1+\lambda}\right)}$ for $i \in \{1, \ldots, d\}$ ;
    /* calculate the mass of $c(\lambda)$                                         */
    $s(\lambda) = \sum_{j=1}^d c_j$;
    if $s(\lambda) > 1$ then
        /* Consider next range $[\lambda, \lambda_M]$                              */
        $l_m = \lambda$ ;
    else
        /* Consider next range $[\lambda_m, \lambda]$                              */
        $\lambda_M = \lambda$ ;
    end
end
$\lambda = \frac{\lambda_m + \lambda_M}{2}$ ;
$c_j(\lambda) = \frac{a_j}{W_0\left(\frac{a_j}{g_j} e^{1+\lambda}\right)}$ for $i \in \{1, \ldots, d\}$ ;
return $c(\lambda)$;
```

Appendix B. Closed-Form Formula for the Symmetrized Log Det Centroids

Consider a set $\mathcal{P} = \{P_1, \ldots, P_n\}$ of n symmetric positive-definite matrices of the d-dimensional SPD cone $\mathrm{Sym}^{++}(d, \mathbb{R})$ weighted by a vector $w = (w_1, \ldots, w_n) \in \Delta_n^\circ$ such that P_i has weight w_i for $i \in \{1, \ldots, n\}$. The log det divergence [59] is a Bregman divergence induced by the strictly convex and differential generator $F_{\mathrm{ld}}(X) = -\log\det(X)$ on $\mathrm{Sym}^{++}(d, \mathbb{R})$ equipped with the inner product $\langle X, Y \rangle = \mathrm{tr}(XY)$ for $X, Y \in \mathrm{Sym}(d, \mathbb{R})$:

$$D_{\mathrm{ld}}(X : Y) = B_{F_{\mathrm{ld}}}(X : Y) = F(X) - F(Y) - \langle X - Y, \nabla F_{\mathrm{ld}}(Y) \rangle. \tag{A1}$$

Since we have $\nabla F_{\mathrm{ld}}(X) = -\nabla \log\det(X) = -\frac{\nabla \det(X)}{\det(X)} = -(X^{-1})^\top$ (hence $\nabla F_{\mathrm{ld}}(X) = -X^{-1}$ for symmetric matrices), it follows that the log det divergence is

$$\begin{aligned} D_{\mathrm{ld}}(X : Y) &= \log\det(YX^{-1}) + \mathrm{tr}((X - Y)Y^{-1}), \\ &= \mathrm{tr}(XY^{-1}) - \log\det(XY^{-1}) - d, \end{aligned}$$

using the properties that $\det(X)\det(Y) = \det(XY)$, $\det(X^{-1}) = \frac{1}{\det(X)}$, and $\mathrm{tr}(I) = d$ where I denotes the $d \times d$ identity matrix. When $d = 1$, we recover the Itakura–Saito divergence [45] obtained for $F_{\mathrm{IS}}(x) = -\log x$ (Burg negative entropy) with $F'_{\mathrm{IS}}(x) = -\frac{1}{x}$:

$$D_{\mathrm{IS}}(x : y) = B_{F_{\mathrm{IS}}}(x : y) = \frac{x}{y} - \log\frac{x}{y} - 1, \quad x, y > 0.$$

The log det divergence is known in statistics as Stein's loss [37,38] and has been used for estimating covariance matrices. The log det divergence S_{ld} satisfies the following invariance properties:

- Inversion invariance: $S_{\mathrm{ld}}(X^{-1}, Y^{-1}) = S_{\mathrm{ld}}(X, Y)$;
- Congruence invariance: for any invertible matrix $A \in \mathrm{GL}(d)$, we have $S_{\mathrm{ld}}(AXA^\top, AY^{-1}A^\top) = S_{\mathrm{ld}}(X, Y)$.

The Jeffreys' symmetrized log det divergence (SLD) is thus

$$\begin{aligned} S_{\mathrm{ld}}(X, Y) &= D_{\mathrm{ld}}(X : Y) + D_{\mathrm{ld}}(Y : X) = \mathrm{tr}\Big(\big(Y^{-1} - X^{-1})(X - Y\big)\Big), \qquad \text{(A2)} \\ &= \mathrm{tr}\Big(X^{-1}Y + Y^{-1}X - 2I\Big). \qquad \text{(A3)} \end{aligned}$$

When $d = 1$, the SLD corresponds to the COSH distance [60] (COSine Hyperbolic distance, the symmetrized Itakura–Saito divergence) when $d = 1$:

$$D_{\mathrm{COSH}}(x : y) = \left(\frac{y}{x} - \frac{1}{x}\right) = \frac{x}{y} + \frac{y}{x} - 2.$$

Consider a family $\mathcal{N}_\mu = \{p_{\mu,\Sigma_1}, \ldots, p_{\mu,\Sigma_n}\}$ of n multivariate normal distributions centered at the same mean $\mu \in \mathbb{R}^d$ with covariance matrices $\Sigma_1, \ldots, \Sigma_n$. The set of same-mean normal distributions forms an exponential family with the natural parameter $\theta = \Sigma^{-1}$ (the precision matrix) corresponding to the sufficient statistics $t(x) = -\frac{1}{2}xx^\top$, and the log-normalizer $F(\theta) = -\frac{1}{2}\log\det(\theta)$. Thus, the Kullback–Leibler divergence between p_{μ,Σ_i} and p_{μ,Σ_j} corresponds to a log det divergence [16]:

$$D_{\mathrm{KL}}[p_{\mu,\Sigma_i}, p_{\mu,\Sigma_j}] = B_F(\theta_j : \theta_i) = D_{\mathrm{ld}}(\Sigma_j^{-1} : \Sigma_i^{-1}),$$

and therefore the Jeffreys divergence $D_J[p_{\mu,\Sigma_i}, p_{\mu,\Sigma_j}]$ corresponds to the matrix COSH/symetrized log-det divergence:

$$D_J[p_{\mu,\Sigma_i}, p_{\mu,\Sigma_j}] = S_{\mathrm{ld}}(\Sigma_i^{-1}, \Sigma_j^{-1}) = \mathrm{tr}\Big(\big(\Sigma_i^{-1} - \Sigma_j^{-1})(\Sigma_j - \Sigma_i\big)\Big). \qquad \text{(A4)}$$

The left KL centroid corresponds to a right Bregman centroid on the natural parameters (the center of mass of the natural parameters), which corresponds to a weighted matrix harmonic mean on the covariance matrices:

$$C_L^{\mathrm{KL}} = C_R^{B_F} = \left(\sum_{i=1}^n w_i \Sigma_i^{-1}\right)^{-1}.$$

The right KL centroid is a left Bregman centroid (i.e., a quasi-arithmetic mean for $h(X) = -X^{-1}$ with $h^{-1}(Y) = -Y^{-1}$) which corresponds to the inverse of the weighted arithmetic mean on the covariance matrices:

$$C_R^{\mathrm{KL}} = C_L^{B_F} = \left(\sum_{i=1}^n w_i \Sigma_i\right)^{-1}.$$

We state the remarkable case of the closed-form formula for the symmetrized Bregman logdet centroid:

Proposition A1 ([29])**.** *The symmetrized log det centroid of a set $\mathcal{P}^w = \{(w_i, P_i)\}$ of n weighted positive-definite matrices is $A\#H$ where $A = \sum_i w_i P_i$ and $H = \left(\sum_i w_i P_i^{-1}\right)^{-1}$ are the weighted arithmetic and harmonic means and $A\#B$ is the matrix geometric mean.*

Since the proof was only briefly sketched in [29], we report a full-length proof for the sake of completeness:

Proof. We have

$$\min_X \sum_i w_i S_{\mathrm{ld}}(X, P_i) \equiv \min_X \mathrm{tr}\left(X^{-1}A + H^{-1}X\right).$$

Setting the gradient of the right-hand-side term to zero using matrix calculus [61] yields

$$\nabla_X \mathrm{tr}\left(X^{-1}A + H^{-1}X\right) = \mathrm{tr}\left(\nabla_X(X^{-1}A + H^{-1}X)\right) = 0.$$

Using the matrix calculus property that $\nabla(M^{-1}) = -M^{-1}(\nabla M)M^{-1}$ for $M = X^{-1}A$, we obtain

$$X^{-1}AX^{-1} - H^{-1} = 0.$$

That is, we need to solve the following Ricatti equation:

$$X^{-1}AX^{-1} = H^{-1}.$$

The well-known Ricatti equation $XA^{-1}X = B$ solves [40] as $X = A\#B$, and therefore we obtain

$$X^{-1} = A^{-1}\#H^{-1}.$$

Finally, we use the invariance property of the geometric mean under matrix inversion, $A^{-1}\#H^{-1} = A\#H$, to obtain the result $C_S^{\mathrm{ld}} = A\#H$. □

The Riemannian Hessian metric $g(\theta)$ induced by $F(\theta) = -\frac{1}{2}\log\det(\theta)$ is

$$g_\theta(S_1, S_2) = \mathrm{tr}\left(\theta^{-1}S_1\theta^{-1}S_2\right),$$

where S_1 and S_2 are two symmetric matrices of the tangent space T_θ at θ. The metric tensor g is commonly called the trace metric or Affine-Invariant Riemannian Metric (AIRM) [62].

It follows that the Riemannian geodesic midpoint is the matrix geometric mean [63] given by

$$X\#Y = X^{\frac{1}{2}}\,(X^{-\frac{1}{2}}\,Y\,X^{-\frac{1}{2}})^{\frac{1}{2}}\,X^{\frac{1}{2}}.$$

We have $\rho(X, X\#Y) = \rho(X\#Y, Y)$, where $\rho(\cdot,\cdot)$ denotes the geodesic length distance on the Riemannian manifold. The geodesic length is given by the following formula [64,65]:

$$\rho(P_1, P_2) = \left\|\log\left(P_1^{-\frac{1}{2}}\,P_2\,P_1^{-\frac{1}{2}}\right)\right\|_F = \sqrt{\sum_{i=1}^d \log^2\lambda_i\left(P_1^{-\frac{1}{2}}\,P_2\,P_1^{-\frac{1}{2}}\right)},$$

where the $\lambda_i(X)$s are the generalized eigenvalues of X.

We state the theorem geometrically characterizing the Jeffreys centroid of a weighted set of centered multivariate normal distributions.

Theorem A1 (The Jeffreys centroid of n weighted centered multivariate normal distributions)**.** *The Jeffreys centroid C_S of a weighted set $\{p_{\mu,\Sigma_i}\}$ of centered normal distributions $N(\mu, \Sigma_i)$ with a weighted $w \in \Delta_n$ corresponds to the midpoint of the Fisher–Rao geodesic linking the left and right SKL centroids:*

$$C_S = \left(\sum_{i=1}^n w_i\Sigma_i\right)\#\left(\sum_{i=1}^n w_i\Sigma_i^{-1}\right)^{-1}, \tag{A5}$$

where $X\#Y$ is the geometric matrix mean:

$$X\#Y = X^{\frac{1}{2}}\,(X^{-\frac{1}{2}}\,Y\,X^{-\frac{1}{2}})^{\frac{1}{2}}\,X^{\frac{1}{2}}.$$

This result first appeared in [29] (Lemma 17.4.3, item 3) and also appeared in an indirect but more general form in [66] (Theorem 5.3). Indeed, in [66], the authors define the regularized symmetric log det divergence as follows:

$$S^{\epsilon}_{\mathrm{ld}}(X,Y) = \mathrm{tr}\Big((X-Y)\Big((Y+\epsilon I)^{-1} - (X+\epsilon I)^{-1}\Big)\Big), \quad \epsilon > 0.$$

This extended definition of the symmetrized logdet divergence allows one to consider degenerate semi-positive definite matrices.

Appendix C. Fisher–Rao Midpoint for Multivariate Normal Distributions

The expression of the Fisher–Rao geodesics for multivariate normal distributions (MVNs) passing through two given d-variate normal distributions was elucidated in [32]. We give below the method for finding those Fisher–Rao MVN midpoints without the underlying geometric explanation that relies on a Riemannian submersion in dimension $2d+1$ [32]. The Python software library `pyBregMan` [52] provides an implementation of those Fisher–Rao MVN midpoints.

Fisher–Rao geodesic midpoint $N = N(\mu, \Sigma)$ of $N_0 = N(\mu_0, \Sigma_0)$ and $N_1 = N(\mu_1, \Sigma_1)$

- For $i \in \{0,1\}$, let $G_0 = M_0\,D_0\,M_0^\top$ and $G_1 = M_1\,D_1\,M_1^\top$, where

$$\begin{aligned} D_0 &= \begin{bmatrix} \Sigma_0^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \Sigma_0 \end{bmatrix}, \\ M_0 &= \begin{bmatrix} I_d & 0 & 0 \\ \mu_0^\top & 1 & 0 \\ 0 & -\mu_0 & I_d \end{bmatrix}, \\ D_1 &= \begin{bmatrix} \Sigma_1^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \Sigma_1 \end{bmatrix}, \\ M_1 &= \begin{bmatrix} I_d & 0 & 0 \\ \mu_1^\top & 1 & 0 \\ 0 & -\mu_1 & I_d \end{bmatrix}, \end{aligned}$$

  where I_d denotes the identity matrix of shape $d \times d$. That is, matrices G_0 and $G_1 \in \mathrm{Sym}_+(2d+1, \mathbb{R})$ can be expressed by block Cholesky factorizations.
- Consider the Riemannian geodesic midpoint G in $\mathrm{Sym}_+(2d+1, \mathbb{R})$ with respect to the trace metric:

$$G = G_0^{\frac{1}{2}} \left(G_0^{-\frac{1}{2}} G_1 G_0^{-\frac{1}{2}}\right)^{\frac{1}{2}} G_0^{\frac{1}{2}}.$$

  In order to compute the matrix power G^p for $p \in \mathbb{R}$, we first calculate the Singular Value Decomposition (SVD) of G: $G = O\,L\,O^\top$ (where O is an orthogonal matrix and $L = \mathrm{diag}(\lambda_1, \ldots, \lambda_{2d+1})$ a diagonal matrix) and then obtain the matrix power as $G^p = O\,L^p\,O^\top$ with $L^p = \mathrm{diag}(\lambda_1^p, \ldots, \lambda_{2d+1}^p)$.

- Retrieve $N = N(\mu, \Sigma)$ from matrix G:

$$\begin{aligned} \Sigma &= [G]_{1:d,1:d}^{-1}, \\ \mu &= \Sigma\,[G]_{1:d,d+1}, \end{aligned}$$

where $[G]_{1:d,1:d}$ denotes the block matrix with rows and columns ranging from one to d extracted from the $(2d+1) \times (2d+1)$ matrix G, and $[G]_{1:d,d+1}$ is similarly the column vector of $\mathbb{R}^d$ extracted from G.

References

1. Jeffreys, H. *The Theory of Probability*; OUP Oxford: Oxford, UK, 1998.
2. Ben-Tal, A.; Charnes, A.; Teboulle, M. Entropic means. *J. Math. Anal. Appl.* **1989**, *139*, 537–551.
3. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466.
4. Amari, S.I. Integration of stochastic models by minimizing α-divergence. *Neural Comput.* **2007**, *19*, 2780–2796.
5. Nielsen, F. On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid. *Entropy* **2020**, *22*, 221.
6. Veldhuis, R. The centroid of the symmetrical Kullback-Leibler distance. *IEEE Signal Process. Lett.* **2002**, *9*, 96–99.
7. Nielsen, F. On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy* **2019**, *21*, 485.
8. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
9. Johnson, D.H.; Sinanovic, S. Symmetrizing the Kullback-Leibler distance. *IEEE Trans. Inf. Theory* **2001**, *1*, 1–10.
10. Fuglede, B.; Topsoe, F. Jensen-Shannon divergence and Hilbert space embedding. In *International Symposium on Information Theory (ISIT)*; IEEE: Piscataway, NJ, USA, 2004; p. 31.
11. Sra, S. Metrics induced by Jensen-Shannon and related divergences on positive definite matrices. *Linear Algebra Its Appl.* **2021**, *616*, 125–138.
12. Vajda, I. On metric divergences of probability measures. *Kybernetika* **2009**, *45*, 885–900.
13. Barndorff-Nielsen, O. *Information and Exponential Families: In Statistical Theory*; John Wiley & Sons: New York, NY, USA, 2014.
14. Amari, S.I. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Berlin/Heidelberg, Germany, 2016.
15. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137.
16. Davis, J.; Dhillon, I. Differential entropic clustering of multivariate gaussians. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 337–344.
17. Murtagh, F.; Legendre, P. Ward's hierarchical agglomerative clustering method: Which algorithms implement Ward's criterion? *J. Classif.* **2014**, *31*, 274–295.
18. Julier, S.; Uhlmann, J.K. General decentralized data fusion with covariance intersection. In *Handbook of Multisensor Data Fusion*; CRC Press: Boca Raton, FL, USA, 2017; pp. 339–364.
19. Liu, Q.; Ihler, A.T. Distributed estimation, information loss and exponential families. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
20. Basseville, M. Divergence measures for statistical data processing: An annotated bibliography. *Signal Process.* **2013**, *93*, 621–633.
21. Chandrasekhar, V.; Takacs, G.; Chen, D.M.; Tsai, S.S.; Reznik, Y.; Grzeszczuk, R.; Girod, B. Compressed histogram of gradients: A low-bitrate descriptor. *Int. J. Comput. Vis.* **2012**, *96*, 384–399.
22. Seal, A.; Karlekar, A.; Krejcar, O.; Gonzalo-Martin, C. Fuzzy *c*-means clustering using Jeffreys-divergence based similarity measure. *Appl. Soft Comput.* **2020**, *88*, 106016.
23. Vasconcelos, N. On the efficient evaluation of probabilistic similarity functions for image retrieval. *IEEE Trans. Inf. Theory* **2004**, *50*, 1482–1496.
24. Ge, P.; Chen, Y.; Wang, G.; Weng, G. An active contour model driven by adaptive local pre-fitting energy function based on Jeffreys divergence for image segmentation. *Expert Syst. Appl.* **2022**, *210*, 118493.
25. Tabibian, S.; Akbari, A.; Nasersharif, B. Speech enhancement using a wavelet thresholding method based on symmetric Kullback–Leibler divergence. *Signal Process.* **2015**, *106*, 184–197.
26. Zhao, Q.; Zhou, G.; Zhang, L.; Cichocki, A. Tensor-variate Gaussian processes regression and its application to video surveillance. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 5–9 May 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 1265–1269.
27. Welk, M.; Feddern, C.; Burgeth, B.; Weickert, J. Tensor median filtering and *M*-smoothing. In *Visualization and Processing of Tensor Fields*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 345–356.
28. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904.
29. Moakher, M.; Batchelor, P.G. Symmetric positive-definite matrices: From geometry to applications and visualization. In *Visualization and Processing of Tensor Fields*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 285–298.
30. Sturm, K.T. Probability measures on metric spaces of nonpositive. *Heat Kernels Anal. Manifolds, Graphs, Metr. Spaces* **2003**, *338*, 357.
31. Nielsen, F. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Process. Lett.* **2013**, *20*, 657–660.
32. Kobayashi, S. Geodesics of multivariate normal distributions and a Toda lattice type Lax pair. *Phys. Scr.* **2023**, *98*, 115241.

33. Corless, R.M.; Gonnet, G.H.; Hare, D.E.; Jeffrey, D.J.; Knuth, D.E. On the Lambert W function. *Adv. Comput. Math.* **1996**, *5*, 329–359.
34. Rockafellar, R.T. Conjugates and Legendre transforms of convex functions. *Can. J. Math.* **1967**, *19*, 200–205.
35. Bullen, P.S.; Bullen, P. Quasi-arithmetic means. In *Handbook of Means and Their Inequalities*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 266–320.
36. Nock, R.; Luosto, P.; Kivinen, J. Mixed Bregman clustering with approximation guarantees. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Antwerp, Belgium, 15–19 September 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 154–169.
37. James, W.; Stein, C. Estimation with quadratic loss. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Los Angeles, CA, USA, 20 June–30 July 1960; Volume 1, pp. 361–379.
38. Salehian, H.; Cheng, G.; Vemuri, B.C.; Ho, J. Recursive estimation of the Stein center of SPD matrices and its applications. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 1793–1800.
39. Skovgaard, L.T. A Riemannian geometry of the multivariate normal model. *Scand. J. Stat.* **1984**, *11*, 211–223.
40. Bhatia, R. The Riemannian mean of positive matrices. In *Matrix Information Geometry*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 35–51.
41. Nakamura, Y. Algorithms associated with arithmetic, geometric and harmonic means and integrable systems. *J. Comput. Appl. Math.* **2001**, *131*, 161–174.
42. Nielsen, F. What is... an Inductive Mean? *Not. Am. Math. Soc.* **2023**, *70*, 1851–1855.
43. Lehmer, D.H. On the compounding of certain means. *J. Math. Anal. Appl.* **1971**, *36*, 183–200.
44. Almkvist, G.; Berndt, B. Gauss, Landen, Ramanujan, the arithmetic-geometric mean, ellipses, π, and the Ladies Diary. *Am. Math. Mon.* **1988**, *95*, 585–608.
45. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J.; Lafferty, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
46. Miyamoto, H.K.; Meneghetti, F.C.; Pinele, J.; Costa, S.I. On closed-form expressions for the Fisher–Rao distance. *Inf. Geom.* **2024**, 1–44. https://doi.org/10.1007/s41884-024-00143-2.
47. Nielsen, F. Approximation and bounding techniques for the Fisher-Rao distances between parametric statistical models; Handbook of Statistics; Elsevier: Amsterdam, The Netherlands, 2024. https://doi.org/10.1016/bs.host.2024.06.003.
48. Karcher, H. Riemannian center of mass and mollifier smoothing. *Commun. Pure Appl. Math.* **1977**, *30*, 509–541.
49. Čencov, N.N. Algebraic foundation of mathematical statistics. *Stat. A J. Theor. Appl. Stat.* **1978**, *9*, 267–276.
50. Calvo, M.; Oller, J.M. A distance between multivariate normal distributions based in an embedding into the Siegel group. *J. Multivar. Anal.* **1990**, *35*, 223–242.
51. Nielsen, F. A simple approximation method for the Fisher–Rao distance between multivariate normal distributions. *Entropy* **2023**, *25*, 654.
52. Nielsen, F.; Soen, A. `pyBregMan`: A Python library for Bregman Manifolds. *arXiv* **2024**, arXiv:2408.04175.
53. Nielsen, F. Divergences Induced by the Cumulant and Partition Functions of Exponential Families and Their Deformations Induced by Comparative Convexity. *Entropy* **2024**, *26*, 193.
54. Shima, H. *The geometry of Hessian Structures*; World Scientific: Singapore, 2007.
55. De Carvalho, M. Mean, what do you Mean? *Am. Stat.* **2016**, *70*, 270–274.
56. Bullen, P.S. *Handbook of Means and Their Inequalities*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013; Volume 560.
57. Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Stat.* **1952**, *23*, 493–507.
58. Nielsen, F. Revisiting Chernoff information with likelihood ratio exponential families. *Entropy* **2022**, *24*, 1400.
59. Kulis, B.; Sustik, M.; Dhillon, I. Learning low-rank kernel matrices. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 505–512.
60. Gray, A.; Markel, J. Distance measures for speech processing. *IEEE Trans. Acoust. Speech, Signal Process.* **1976**, *24*, 380–391.
61. Petersen, K.B.; Pedersen, M.S. The matrix cookbook. *Tech. Univ. Den.* **2008**, *7*, 510.
62. Thanwerdas, Y.; Pennec, X. $O(n)$-invariant Riemannian metrics on SPD matrices. *Linear Algebra Its Appl.* **2023**, *661*, 163–201.
63. Bhatia, R.; Holbrook, J. Riemannian geometry and matrix geometric means. *Linear Algebra Its Appl.* **2006**, *413*, 594–618.
64. Siegel, C. Symplectic geometry. *Am. J. Math.* **1964**, *65*, 1–86.
65. James, A.T. The variance information manifold and the functions on it. In *Multivariate Analysis–III*; Elsevier: Amsterdam, The Netherlands, 1973; pp. 157–169.
66. Kim, S.; Lawson, J.; Lim, Y. The matrix geometric mean of parameterized, weighted arithmetic and harmonic means. *Linear Algebra Its Appl.* **2011**, *435*, 2114–2131.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.