entropy

MDPI

Article

Fast Approximations of the Jeffreys Divergence between Univariate Gaussian Mixtures via Mixture Conversions to Exponential-Polynomial Distributions

Frank Nielsen

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

Abstract: The Jeffreys divergence is a renown arithmetic symmetrization of the oriented Kullback–Leibler divergence broadly used in information sciences. Since the Jeffreys divergence between Gaussian mixture models is not available in closed-form, various techniques with advantages and disadvantages have been proposed in the literature to either estimate, approximate, or lower and upper bound this divergence. In this paper, we propose a simple yet fast heuristic to approximate the Jeffreys divergence between two univariate Gaussian mixtures with arbitrary number of components. Our heuristic relies on converting the mixtures into pairs of dually parameterized probability densities belonging to an exponential-polynomial family. To measure with a closed-form formula the goodness of fit between a Gaussian mixture and an exponential-polynomial density approximating it, we generalize the Hyvärinen divergence to α-Hyvärinen divergences. In particular, the 2-Hyvärinen divergence allows us to perform model selection by choosing the order of the exponential-polynomial densities used to approximate the mixtures. We experimentally demonstrate that our heuristic to approximate the Jeffreys divergence between mixtures improves over the computational time of stochastic Monte Carlo estimations by several orders of magnitude while approximating the Jeffreys divergence reasonably well, especially when the mixtures have a very small number of modes.

Keywords: Gaussian mixture model; Jeffreys divergence; mixture family; exponential-polynomial family; Maximum Likelihood Estimator; Score Matching Estimator; Hyvärinen divergence; relative Fisher information; moment matrix; Hankel matrix

check for updates

Citation: Nielsen, F. Fast Approximations of the Jeffreys Divergence between Univariate Gaussian Mixtures via Mixture Conversions to Exponential-Polynomial Distributions. *Entropy* **2021**, *23*, 1417. https://doi.org/10.3390/e23111417

Academic Editor: Takuya Yamano

Received: 7 September 2021
Accepted: 26 October 2021
Published: 28 October 2021

1. Introduction

1.1. Statistical Mixtures and Statistical Divergences

We consider the problem of approximating the Jeffreys divergence [1] between two finite univariate continuous mixture models [2] $m(x) = \sum_{i=1}^{k} w_i p_i(x)$ and $m'(x) = \sum_{i=1}^{k'} w'_i p'_i(x)$ with continuous component distributions p_i's and p_i''s defined on a coinciding support $\mathcal{X} \subset \mathbb{R}$. The mixtures $m(x)$ and $m'(x)$ may have a different number of components (i.e., $k \neq k'$). Historically, Pearson [3] first considered a univariate Gaussian mixture of two components for modeling the distribution of the ratio of forehead breadth to body length of a thousand crabs in 1894 (Pearson obtained a unimodal mixture).

Although our work applies to any continuous mixtures of an exponential family (e.g., Rayleigh mixtures [4] with restricted support $\mathcal{X} = \mathbb{R}_+$), we explain our method for the most prominent family of mixtures encountered in practice: the Gaussian mixture models or GMMs for short. In the remainder, a univariate GMM $m(x) = \sum_{i=1}^{k} w_i p_{\mu_i,\sigma_i}(x)$ with k Gaussian components

$$p_i(x) = p_{\mu_i,\sigma_i}(x) := \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right),$$

is called a k-GMM.

The Kullback–Leibler divergence (KLD) [5,6] $D_{\mathrm{KL}}[m : m']$ between two mixtures m and m' is:

$$D_{\mathrm{KL}}[m : m'] := \int_{\mathcal{X}} m(x) \log\left(\frac{m(x)}{m'(x)}\right) \mathrm{d}x. \tag{1}$$

The KLD is an oriented divergence since $D_{\mathrm{KL}}[m : m'] \neq D_{\mathrm{KL}}[m' : m]$.

The Jeffreys divergence (JD) [1] $D_J[m, m']$ is the arithmetic symmetrization of the forward KLD and the reverse KLDs:

$$\begin{aligned} D_J[m, m'] &:= D_{\mathrm{KL}}[m : m'] + D_{\mathrm{KL}}[m' : m], & (2)\\ &= \int_{\mathcal{X}} (m(x) - m'(x)) \log\left(\frac{m(x)}{m'(x)}\right) \mathrm{d}x. & (3)\end{aligned}$$

The JD is a symmetric divergence: $D_J[m, m'] = D_J[m', m]$. In the literature, the Jeffreys divergence [7] has also been called the J-divergence [8,9], the symmetric Kullback–Leibler divergence [10] and sometimes the symmetrical Kullback–Leibler divergence [11,12]. In general, it is provably hard to calculate the definite integral of the KLD between two continuous mixtures in closed-form: For example, the KLD between two GMMs has been shown to be non-analytic [13]. Thus, in practice, when calculating the JD between two GMMs, one can either approximate [14,15], estimate [16], or bound [17,18] the KLD between mixtures. Another approach to bypass the computational intractability of calculating the KLD between mixtures consists of designing new types of divergences that admit closed-form expressions for mixtures. See, for example, the Cauchy–Schwarz divergence [19] or the total square divergence [20] (a total Bregman divergence) that admit the closed-form formula when handling GMMs. The total square divergence [20] is invariant to rigid transformations and provably robust to outliers in clustering applications.

In practice, to estimate the KLD between mixtures, one uses the following Monte Carlo (MC) estimator:

$$\hat{D}_{\mathrm{KL}}^{\mathcal{S}_s}[m : m'] := \frac{1}{s} \sum_{i=1}^{s} \left(\log\left(\frac{m(x_i)}{m'(x_i)}\right) + \frac{m'(x_i)}{m(x_i)} - 1 \right) \geq 0,$$

where $\mathcal{S}_s = \{x_1, \ldots, x_s\}$ is s independent and identically distributed (i.i.d.) samples from $m(x)$. This MC estimator is by construction always non-negative and therefore consistent. That is, we have $\lim_{s\to\infty} \hat{D}_{\mathrm{KL}}^{\mathcal{S}_s}[m : m'] = D_{\mathrm{KL}}[m : m']$ under mild conditions [21].

Similarly, we estimate the Jeffreys divergence via MC sampling as follows:

$$\hat{D}_J^{\mathcal{S}_s}[m, m'] := \frac{1}{s} \sum_{i=1}^{s} 2 \frac{(m(x_i) - m'(x_i))}{m(x_i) + m'(x_i)} \log\left(\frac{m(x_i)}{m'(x_i)}\right) \geq 0, \tag{4}$$

where $\mathcal{S}_s = \{x_1, \ldots, x_s\}$ are s i.i.d. samples from the "middle mixture" $m_{12}(x) := \frac{1}{2}(m(x) + m'(x))$. By choosing the middle mixture $m_{12}(x)$ for sampling, we ensure that we keep the symmetric property of the JD (i.e., $\hat{D}_J^{\mathcal{S}_s}[m, m'] = \hat{D}_J^{\mathcal{S}_s}[m', m]$), and we also have consistency under mild conditions [21]: $\lim_{s\to\infty} \hat{D}_J^{\mathcal{S}_s}[m, m'] = D_J[m, m']$. The time complexity to stochastically estimate the JD is $\tilde{O}((k + k')s)$, with s typically ranging from 10^4 to 10^6 in applications. Notice that the number of components of a mixture can be very large (e.g., $k = O(n)$ for n input data when using Kernel Density Estimators [2]). KDEs may also have a large number of components and may potentially exhibit many spurious modes visualized as small bumps when plotting the densities.

1.2. Jeffreys Divergence between Densities of an Exponential Family

We consider approximating the JD by converting continuous mixtures into densities of exponential families [22]. A continuous exponential family (EF) $\mathcal{E}_t$ of order D is defined as a family of probability density functions with support $\mathcal{X}$ and the probability density function:

$$\mathcal{E}_t := \left\{ p_\theta(x) := \exp\left(\sum_{i=1}^{D} \theta_i t_i(x) - F(\theta) \right) \ : \ \theta \in \Theta \right\},$$

where $F(\theta)$ is called the log-normalizer, which ensures the normalization of $p_\theta(x)$ (i.e., $\int_{\mathcal{X}} p_\theta(x) \mathrm{d}x = 1$):

$$F(\theta) = \log\left(\int_{\mathcal{X}} \exp\left(\sum_{i=1}^{D} \theta_i t_i(x) \right) \mathrm{d}x \right).$$

Parameter $\theta \in \Theta \subset \mathbb{R}^D$ is called the natural parameter, and the functions $t_1(x), \ldots$, $t_D(x)$ are called the sufficient statistics [22]. Let Θ denote the natural parameter space: $\Theta := \{\theta \ : \ F(\theta) < \infty\}$, an open convex domain for regular exponential families [22]. The exponential family is said to be minimal when the functions $1, t_1(x), \ldots, t_D(x)$ are linearly independent.

It is well-known that one can bypass the definite integral calculation of the KLD when the probability density functions p_θ and $p_{\theta'}$ belong to the same exponential family [23,24]:

$$D_{\mathrm{KL}}[p_\theta : p_\theta] = B_F(\theta' : \theta),$$

where $B_F(\theta_2 : \theta_1)$ is the Bregman divergence induced by the log-normalizer, a strictly convex real-analytic function [22]. The Bregman divergence [25] between two parameters θ_1 and θ_2 for a strictly convex and smooth generator F is defined by:

$$B_F(\theta_1 : \theta_2) := F(\theta_1) - F(\theta_2) - (\theta_1 - \theta_2)^\top \nabla F(\theta_2). \tag{5}$$

Thus, the Jeffreys divergence between two pdfs p_θ and $p_{\theta'}$ belonging to the same exponential family is a symmetrized Bregman divergence [26]:

$$\begin{aligned} D_J[p_\theta : p_\theta] &= B_F(\theta' : \theta) + B_F(\theta : \theta'), \\ &= (\theta' - \theta)^\top (\nabla F(\theta') - \nabla F(\theta)). \end{aligned}$$

Let $F^*(\eta)$ denote the Legendre–Fenchel convex conjugate of $F(\theta)$:

$$F^*(\eta) := \sup_{\theta \in \Theta} \{\theta^\top \eta - F(\theta)\}. \tag{6}$$

The Legendre transform ensures that $\eta = \nabla F(\theta)$ and $\theta = \nabla F^*(\eta)$, and the Jeffreys divergence between two pdfs p_θ and $p_{\theta'}$ belonging to the same exponential family is:

$$D_J[p_\theta : p_\theta] = (\theta' - \theta)^\top (\eta' - \eta). \tag{7}$$

Notice that the log-normalizer $F(\theta)$ does not appear explicitly in the above formula.

1.3. A Simple Approximation Heuristic

Densities p_θ of an exponential family admit a dual parameterization [22]: $\eta = \eta(\theta) := E_{p_\theta}[t(x)] = \nabla F(\theta)$, called the moment parameterization (or mean parameterization). Let H denote the moment parameter space. Let us use the subscript and superscript notations to emphasize the coordinate system used to index a density: In our notation, we thus write $p_\theta(x) = p^\eta(x)$.

In view of Equation (7), our method to approximate the Jeffreys divergence between mixtures m and m' consists of first converting those mixtures m and m' into pairs of polynomial exponential densities (PEDs) in Section 2. To convert a mixture $m(x)$ into a pair $(p_{\bar{\theta}_1}, p^{\bar{\eta}_2})$ dually parameterized (but not dual because $\bar{\eta}_2 \neq \nabla F(\bar{\theta}_1)$), we shall consider "integral extensions" (or information projections) of the Maximum Likelihood Estimator [22] (MLE estimates in the moment parameter space $H = \{\nabla F(\theta) \ : \ \theta \in \Theta\}$)

and of the Score Matching Estimator [27] (SME estimates in the natural parameter space $\Theta = \{\nabla F^*(\eta) \,:\, \eta \in H\}$).

We shall consider polynomial exponential families [28] (PEFs) also called exponential-polynomial families (EPFs) [29]. PEFs $\mathcal{E}_D$ are regular minimal exponential families with polynomial sufficient statistics $t_i(x) = x^i$ for $i \in \{1, \ldots, D\}$. For example, the exponential distributions $\{p_\lambda(x) = \lambda \exp(-\lambda x)\}$ form a PEF with $D = 1$, $t(x) = x$ and $\mathcal{X} = \mathbb{R}_+$, and the normal distributions form an EPF with $D = 2$, $t(x) = [x\ x^2]^\top$ and $\mathcal{X} = \mathbb{R}$, etc. Although the log-normalizer $F(\theta)$ can be obtained in closed-form for lower order PEFs (e.g., $D = 1$ or $D = 2$) or very special subfamilies (e.g., when $D = 1$ and $t_1(x) = x^k$, exponential-monomial families [30]), a no-closed form formula is available for $F(\theta)$ of EPFs in general as soon $D \geq 4$ [31,32], and the cumulant function $F(\theta)$ is said to be computationally intractable. Notice that when $\mathcal{X} = \mathbb{R}$, the leading coefficient θ_D is negative for even integer order D. EPFs are attractive because these families can universally model any smooth multimodal distribution [28] and require fewer parameters in comparison to GMMs: Indeed, a univariate k-GMM $m(x)$ (at most k modes and $k-1$ antimodes) requires $3k-1$ parameters to specify $m(x)$ (or $k+1$ for a KDE with constant kernel width σ or $2k-1$ for a KDE with varying kernel widths, but then $k = n$ observations). A density of an EPF of order D is called an exponential-polynomial density (EPD) and requires D parameters to specify θ, with, at most, $\frac{D}{2}$ modes (and $\frac{D}{2} - 1$ antimodes). The case of the quartic (polynomial) exponential densities $\mathcal{E}_4$ ($D = 4$) has been extensively investigated in [31,33–37]. Armstrong and Brigo [38] discussed order-6 PEDs, and Efron and Hastie reported and order-7 PEF in their textbook (see Figure 5.7 of [39]). Figure 1 displays two examples of converting a GMM into a pair of dually parameterized exponential-polynomial densities.

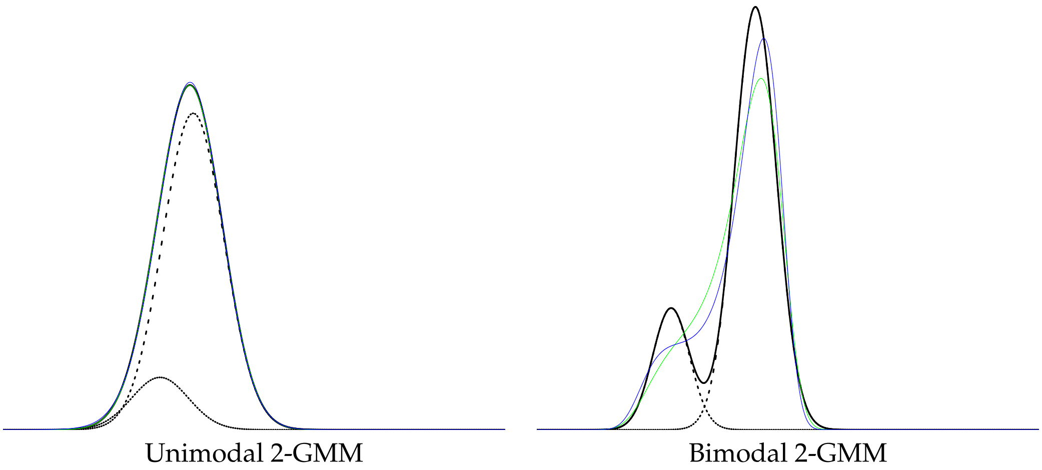


Figure 1. Two examples illustrating the conversion of a GMM m (black) of $k = 2$ components (dashed black) into a pair of polynomial exponential densities of order $D = 4$ $(p_{\bar{\theta}_{\mathrm{SME}}}, p^{\bar{\eta}_{\mathrm{MLE}}})$. PED $p_{\bar{\theta}_{\mathrm{SME}}}$ is displayed in green, and PED $p^{\bar{\eta}_{\mathrm{MLE}}}$ is displayed in blue. To display $p^{\bar{\eta}_{\mathrm{MLE}}}$, we first converted $\bar{\eta}_{\mathrm{MLE}}$ to $\tilde{\bar{\theta}}_{\mathrm{MLE}}$ using an iterative linear system descent method (ILSDM), and we numerically estimated the normalizing factors $Z(\bar{\theta}_{\mathrm{SME}})$ and $Z(\bar{\eta}_{\mathrm{MLE}})$ to display the normalized PEDs.

Then by converting both mixture m and mixture m' into pairs of dually natural/moment parameterized unnormalized PEDs, i.e., $m \to (q_{\bar{\theta}_{\mathrm{SME}}}, q_{\bar{\eta}_{\mathrm{MLE}}})$ and $m' \to (q_{\bar{\theta}'_{\mathrm{SME}}}, q'_{\bar{\eta}_{\mathrm{MLE}}})$, we approximate the JD between mixtures m and m' by using the four parameters of the PEDs

$$D_J[m, m'] \approx (\bar{\theta}'_{\mathrm{SME}} - \bar{\theta}_{\mathrm{SME}})^\top (\bar{\eta}'_{\mathrm{MLE}} - \bar{\eta}_{\mathrm{MLE}}). \tag{8}$$

Let Δ_J denote the approximation formula obtained from the two pairs of PEDs:

$$\Delta_J[p_{\theta_{\mathrm{SME}}}, p^{\eta_{\mathrm{MLE}}}; p_{\theta'_{\mathrm{SME}}}, p^{\eta'_{\mathrm{MLE}}}] := (\theta'_{\mathrm{SME}} - \theta_{\mathrm{SME}})^\top (\eta'_{\mathrm{MLE}} - \eta_{\mathrm{MLE}}). \tag{9}$$

Let $\Delta_J(\theta_{\mathrm{SME}}, \eta_{\mathrm{MLE}}; \theta'_{\mathrm{SME}}, \eta'_{\mathrm{MLE}}) := \Delta_J[p_{\theta_{\mathrm{SME}}}, p^{\eta_{\mathrm{MLE}}}; p_{\theta'_{\mathrm{SME}}}, p^{\eta'_{\mathrm{MLE}}}]$. Then we have

$$D_J[m, m'] \approx \tilde{D}_J[m, m'] := \Delta_J(\theta_{\mathrm{SME}}, \eta_{\mathrm{MLE}}; \theta'_{\mathrm{SME}}, \eta'_{\mathrm{MLE}}).$$

Note that Δ_J is not a proper divergence as it may be negative since, in general, $\bar{\eta}_{\mathrm{MLE}} \neq \nabla F(\bar{\theta}_{\mathrm{SME}})$. That is, Δ_J may not satisfy the law of the indiscernibles. Approximation Δ_J is exact when $k_1 = k_2 = 1$, with both m and m' belonging to an exponential family.

We experimentally show in Section 4 that the $\tilde{D}_J$ heuristic yields fast approximations of the JD compared to the MC baseline estimations by several order of magnitudes while approximating the JD reasonably well when the mixtures have a small number of modes.

For example, Figure 2 displays the unnormalized PEDs obtained for two Gaussian mixture models ($k_1 = 10$ components and $k_2 = 11$ components) into PEDs of a PEF of order $D = 8$. The MC estimation of the JD with $s = 10^6$ samples yields $0.2633\ldots$, while the PED approximation of Equation (8) on corresponding PEFs yields $0.2618\ldots$ (the relative error is $0.00585\ldots$ or about $0.585\ldots\%$). It took about 2642.581 milliseconds (with $s = 10^6$ on a Dell Inspiron 7472 laptop) to MC estimate the JD, while it took about 0.827 milliseconds with the PEF approximation. Thus, we obtained a speed-up factor of about 3190 (three orders of magnitude) for this particular example. Notice that when viewing Figure 2, we tend to visually evaluate the dissimilarity using the total variation distance (a metric distance):

$$D_{\mathrm{TV}}[m, m'] := \frac{1}{2} \int |m(x) - m'(x)| \mathrm{d}x,$$

rather than by a dissimilarity relating to the KLD. Using Pinsker's inequality [40,41], we have $D_J[m, m'] \geq D_{\mathrm{TV}}[m, m']^2$ and $D_{\mathrm{TV}}[m, m'] \in [0, 1]$. Thus, large TV distance (e.g., $D_{\mathrm{TV}}[m, m'] = 0.1$) between mixtures may have a small JD since Pinsker's inequality yields $D_J[m, m'] \geq 0.01$.

Let us point out that our approximation heuristic is deterministic, while the MC estimations are stochastic: That is, each MC run (Equation (4)) returns a different result, and a single MC run may yield a very bad approximation of the true Jeffreys divergence.

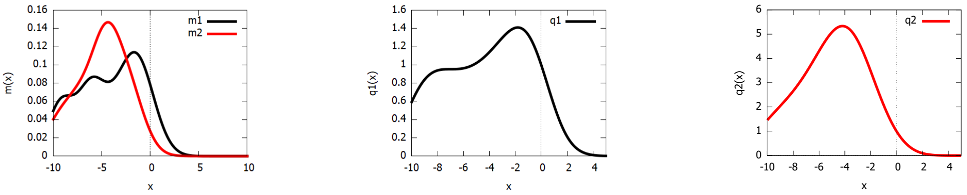


Figure 2. Two mixtures m_1 (black) and m_2 (red) of $k_1 = 10$ components and $k_2 = 11$ components (**left**), respectively. The unnormalized PEFs $q_{\bar{\theta}_1} = \tilde{p}_{\bar{\theta}_1}$ (**middle**) and $q_{\bar{\theta}_2} = \tilde{p}_{\bar{\theta}_2}$ (**right**) of order $D = 8$. Jeffreys divergence (about 0.2634) is approximated using PEDs within 0.6% compared to the Monte Carlo estimate with a speed factor of about 3190. Notice that displaying $p_{\bar{\theta}_1}$ and $p_{\bar{\theta}_2}$ on the same PDF canvas as the mixtures would require calculating the partition functions $Z(\bar{\theta}_1)$ and $Z(\bar{\theta}_2)$ (which we do not in this figure). The PEDs $q^{\bar{\eta}_1}$ and $q^{\bar{\eta}_2}$ of the pairs $(\bar{\theta}_1, \bar{\eta}_1)$ and $(\bar{\theta}_2, \bar{\eta}_2)$ parameterized in the moment space are not shown here.

We compare our fast heuristic $\tilde{D}_J[m, m'] = (\theta'_{\mathrm{SME}} - \theta_{\mathrm{SME}})^\top (\eta'_{\mathrm{MLE}} - \eta_{\mathrm{MLE}})$ with two more costly methods relying on numerical procedures to convert natural ↔ moment parameters:

1. Simplify GMMs m_i into $p^{\bar{\eta}_i^{\mathrm{MLE}}}$, and approximately convert the $\bar{\eta}_i^{\mathrm{MLE}}$'s into $\tilde{\theta}_i^{\mathrm{MLE}}$'s. Then approximate the Jeffreys divergence as

$$D_J[m_1, m_2] \simeq \tilde{\Delta}_J^{\mathrm{MLE}}[m_1, m_2] := (\tilde{\theta}_2^{\mathrm{MLE}} - \tilde{\theta}_1^{\mathrm{MLE}})^\top (\bar{\eta}_2^{\mathrm{MLE}} - \bar{\eta}_1^{\mathrm{MLE}}). \qquad (10)$$

2. Simplify GMMs m_i into $p_{\bar{\theta}_i^{\mathrm{SME}}}$, and approximately convert the $\bar{\theta}_i^{\mathrm{SME}}$'s into $\tilde{\eta}_i^{\mathrm{SME}}$'s. Then approximate the Jeffreys divergence as

$$D_J[m_1, m_2] \simeq \tilde{\Delta}_J^{\mathrm{SME}}(m_1, m_2) = (\bar{\theta}_2^{\mathrm{SME}} - \bar{\theta}_1^{\mathrm{SME}})^\top (\tilde{\eta}_2^{\mathrm{SME}} - \tilde{\eta}_1^{\mathrm{SME}}). \tag{11}$$

1.4. Contributions and Paper Outline

Our contributions are summarized as follows:

- We explain how to convert any continuous density $r(x)$ (including GMMs) into a polynomial exponential density in Section 2 using integral-based extensions of the Maximum Likelihood Estimator [22] (MLE estimates in the moment parameter space H, Theorem 1 and Corollary 1) and the Score Matching Estimator [27] (SME estimates in the natural parameter space Θ, Theorem 3). We show a connection between SME and the Moment Linear System Estimator [28] (MLSE).
- We report a closed-form formula to evaluate the goodness-of-fit of a polynomial family density to a GMM in Section 3 using an extension of the Hyvärinen divergence [42] (Theorem 4) and discuss the problem of model selection for choosing the order D of the polynomial exponential family.
- We show how to approximate the Jeffreys divergence between GMMs using a pair of natural/moment parameter PED conversion and present experimental results that display a gain of several orders of magnitude of performance when compared to the vanilla Monte Carlo estimator in Section 4. We observe that the quality of the approximations depend on the number of modes of the GMMs [43]. However, calculating or counting the modes of a GMM is a difficult problem in its own [43].

The paper is organized as follows: In Section 2, we show how to convert arbitrary probability density functions into polynomial exponential densities using the integral-based Maximum Likelihood Estimator (MLE) and Score Matching Estimator (SME). We describe a Maximum Entropy method to iteratively convert moment parameters into natural parameters in Section 2.3.1. It is followed by Section 3, which shows how to calculate in closed-form the order-2 Hyvärinen divergence between a GMM and a polynomial exponential density. We use this criterion to perform model selection. Section 4 presents our computational experiments that demonstrate a gain of several orders of magnitudes for GMMs with a small number of modes. Finally, we conclude in Section 5.

2. Converting Finite Mixtures to Exponential Family Densities

We report two generic methods to convert a mixture $m(x)$ into a density $p_\theta(x)$ of an exponential family: The first method extending the MLE in Section 2.1 proceeds using the mean parameterization η, while the second method extending the SME in Section 2.2 uses the natural parameterization of the exponential family. We then describe how to convert the moments parameters into natural parameters (and vice versa) for polynomial exponential families in Section 2.3. We show how to instantiate these generic conversion methods for GMMs: It requires calculating non-central moments of GMMs in closed-form. The efficient computations of raw moments of GMMs is detailed in Section 2.4.

2.1. Conversion Using the Moment Parameterization (MLE)

Let us recall that in order to estimate the moment or mean parameter $\hat{\eta}_{\mathrm{MLE}}$ of a density belonging an exponential family

$$\mathcal{E}_t := \left\{ p_\theta(x) = \exp\left(t(x)^\top \theta - F(\theta) \right) \right\}$$

with a sufficient statistic vector $t(x) = [t_1(x)\ \ldots\ t_D(x)]^\top$ from an i.i.d. sample set $x_1, \ldots, x_n$, the Maximum Likelihood Estimator (MLE) [22,44] yields

$$\begin{aligned} & \max_\theta \prod_{i=1}^n p_\theta(x_i), && (12)\\ \equiv\ & \max_\theta \sum_{i=1}^n \log p_\theta(x_i), && (13)\\ =\ & \max_\theta E(\theta) := \left(\sum_{i=1}^n t(x_i)^\top \theta\right) - nF(\theta), && (14)\\ \Rightarrow\ & \hat{\eta}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^n t(x_i). && (15)\end{aligned}$$

In statistics, Equation (14) is called the estimating equation. The MLE exists under mild conditions [22] and is unique since the Hessian $\nabla^2 E(\theta) = \nabla^2 F(\theta)$ of the estimating equation is positive-definite (log-normalizers $F(\theta)$ are always strictly convex and real analytic [22]). The MLE is consistent and asymptotically normally distributed [22]. Furthermore, since the MLE satisfies the equivariance property [22], we have $\hat{\theta}_{\mathrm{MLE}} = \nabla F^*(\hat{\eta}_{\mathrm{MLE}})$, where ∇F^* denotes the gradient of the conjugate function $F^*(\eta)$ of the cumulant function $F(\theta)$ of the exponential family. In general, ∇F^* is intractable for PEDs with $D \geq 4$.

By considering the empirical distribution

$$p_e(x) := \frac{1}{n}\sum_{i=1}^s \delta_{x_i}(x),$$

where $\delta_{x_i}(\cdot)$ denotes the Dirac distribution at location x_i, we can formulate the MLE problem as a minimum KLD problem between the empirical distribution and a density of the exponential family:

$$\begin{aligned}\min_\theta D_{\mathrm{KL}}[p_e : p_\theta] &= \min -H[p_e] - E_{p_e}[\log p_\theta(x)],\\ &\equiv \max_\theta \frac{1}{n}\sum_{i=1}^n \log p_\theta(x_i),\end{aligned}$$

since the entropy term $H[p_e]$ is independent of θ.

Thus, to convert an arbitrary smooth density $r(x)$ into a density p_θ of an exponential family $\mathcal{E}_t$, we have to solve the following minimization problem:

$$\min_{\theta\in\Theta} D_{\mathrm{KL}}[r : p_\theta].$$

Rewriting the minimization problem as:

$$\begin{aligned} & \min_\theta D_{\mathrm{KL}}[r : p_\theta] = -\int r(x)\log p_\theta(x)\mathrm{d}x + \int r(x)\log r(x)\mathrm{d}x,\\ \equiv\ & \min_\theta -\int r(x)\log p_\theta(x)\mathrm{d}x,\\ =\ & \min_\theta \int r(x)(F(\theta) - \theta^\top t(x))\mathrm{d}x,\\ =\ & \min_\theta \bar{E}(\theta) = F(\theta) - \theta^\top E_r[t(x)],\end{aligned}$$

we obtain

$$\bar{\eta}_{\mathrm{MLE}}(r) := E_r[t(x)] = \int_{\mathcal{X}} r(x)t(x)\mathrm{d}x. \qquad (16)$$

The minimum is unique since $\nabla^2\bar{E}(\theta) = \nabla^2 F(\theta) \succ 0$ (positive-definite matrix). This conversion procedure $r(x) \to p^{\bar{\eta}_{\mathrm{MLE}}(r)}(x)$ can be interpreted as an integral extension of the MLE, hence the $\bar{\cdot}$ notation in $\bar{\eta}_{\mathrm{MLE}}$. Notice that the ordinary MLE is $\hat{\eta}_{\mathrm{MLE}} = \bar{\eta}_{\mathrm{MLE}}(p_e)$ obtained for the empirical distribution: $r = p_e$: $\bar{\eta}_{\mathrm{MLE}}(p_e) = \frac{1}{n}\sum_{i=1}^n t(x_i)$.

Theorem 1. *The best density $p^{\bar{\eta}}(x)$ of an exponential family $\mathcal{E}_t = \{p_\theta : \theta \in \Theta\}$ minimizing the Kullback–Leibler divergence $D_{\mathrm{KL}}[r : p_\theta]$ between a density r and a density p_θ of an exponential family $\mathcal{E}_t$ is $\bar{\eta} = E_r[t(x)] = \int_{\mathcal{X}} r(x)t(x)\mathrm{d}x$.*

Notice that when $r = p_\theta$, we obtain $\bar{\eta} = E_{p_\theta}[t(x)] = \eta$ so that the method $\bar{\eta}_{\mathrm{MLE}}(r)$ is consistent (by analogy to the finite i.i.d. MLE case): $\bar{\eta}_{\mathrm{MLE}}(p_\theta) = \eta = \nabla F(\theta)$.

The KLD right-sided minimization problem can be interpreted as an information projection of r onto $\mathcal{E}_t$. As a corollary of Theorem 1, we obtain:

Corollary 1 (Best right-sided KLD simplification of a mixture)**.** *The best right-sided KLD simplification of a homogeneous mixture of exponential families [2] $m(x) = \sum_{i=1}^k w_i p_{\theta_i}(x)$ with $p_{\theta_i} \in \mathcal{E}_t$, i.e., $\min_{\theta\in\Theta} D_{\mathrm{KL}}[m : p_\theta]$, into a single component $p^\eta(x)$ is given by $\eta = \hat{\eta}_{\mathrm{MLE}}(m) = E_m[t(x)] = \sum_{i=1}^k \eta_i = \bar{\eta}$.*

Equation (16) allows us to greatly simplify the proofs reported in [45] for mixture simplifications that involved the explicit use of the Pythagoras' theorem in the dually flat spaces of exponential families [42]. Figure 3 displays the geometric interpretation of the best KLD simplification of a GMM with ambient space the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_L)$, where μ_L denotes the Lebesgue measure and $\mathcal{B}(\mathbb{R})$ the Borel σ-algebra of $\mathbb{R}$.

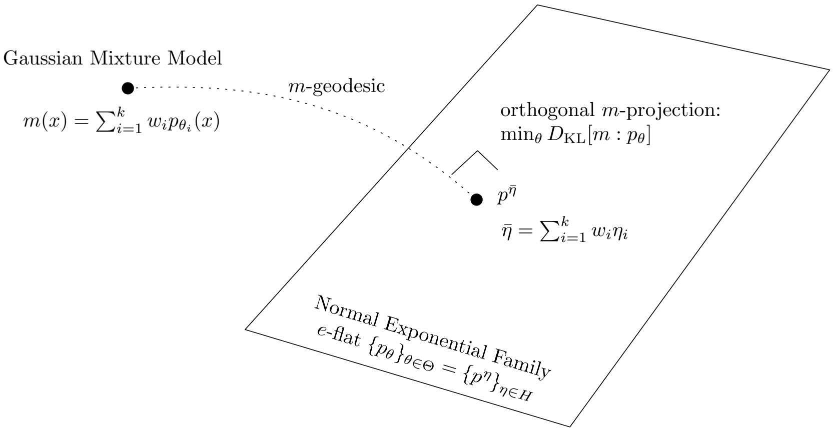

Gaussian Mixture Model
$m(x) = \sum_{i=1}^k w_i p_{\theta_i}(x)$
m-geodesic
orthogonal m-projection:
$\min_\theta D_{\mathrm{KL}}[m : p_\theta]$
$p^{\bar{\eta}}$
$\bar{\eta} = \sum_{i=1}^k w_i \eta_i$
Normal Exponential Family
e-flat $\{p_\theta\}_{\theta\in\Theta} = \{p^\eta\}_{\eta\in H}$

Figure 3. The best simplification of a GMM $m(x)$ into a single normal component p_{θ^*} ($\min_{\theta\in\Theta} D_{\mathrm{KL}}[m : p_\theta] = \min_{\eta\in H} D_{\mathrm{KL}}[m : p^\eta]$) is geometrically interpreted as the unique m-projection of $m(x)$ onto the Gaussian family (a e-flat): We have $\eta^* = \bar{\eta} = \sum_{i=1}^k \eta_i$.

Let us notice that Theorem 1 yields an algebraic system for polynomial exponential densities, i.e., $E_m[x^i] = \bar{\eta}_i$ for $i \in \{1, \ldots, D\}$, to compute $\bar{\eta}_{\mathrm{MLE}}(m)$ for a given GMM $m(x)$ (since raw moments $E_m[x^i]$ are algebraic). In contrast with this result, the MLE of i.i.d. observations is in general not an algebraic function [46] but a transcendental function.

2.2. Converting to a PEF Using the Natural Parameterization (SME)

Integral-Based Score Matching Estimator (SME)

To convert the density $r(x)$ into an exponential density with sufficient statistics $t(x)$, we can also use the Score Matching Estimator [27,47] (SME). The Score Matching Estimator minimizes the Hyvärinen divergence D_H (Equation (4) of [47]):

$$D_H[p : p_\theta] := \frac{1}{2}\int \|\nabla_x \log p(x) - \nabla_x \log p_\theta(x)\|^2\, p(x)\mathrm{d}x.$$

The Hyvärinen divergence is also known as half of the relative Fisher information in the optimal transport community (Equation (8) of [48] or Equation (2.2) in [49]), where it is defined for two measures μ and ν as follows:

$$I[\mu : \nu] := \int_{\mathcal{X}} \left\| \nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \right\|^2 \mathrm{d}\mu = 4 \int_{\mathcal{X}} \left\| \nabla \sqrt{\frac{\mathrm{d}\mu}{\mathrm{d}\nu}} \right\|^2 \mathrm{d}\nu.$$

Moreover, the relative Fisher information can be defined on complete Riemannian manifolds [48].

That is, we convert a density $r(x)$ into an exponential family density $p_\theta(x)$ using the following minimizing problem:

$$\theta_{\mathrm{SME}}(r) = \min_{\theta \in \Theta} D_H[r : p_\theta].$$

Beware that in statistics, the score $s_\theta(x)$ is defined by $\nabla_\theta \log p_\theta(x)$, but in Score Matching, we refer to the "data score" defined by $\nabla_x \log p_\theta(x)$. Hyvärinen [47] gave an explanation of the naming "score" using a spurious location parameter.

- Generic solution: It can be shown that for exponential families [47], we obtain the following solution:

$$\theta_{\mathrm{SME}}(r) = -(E_r[A(x)])^{-1} \times (E_r[b(x)]), \tag{17}$$

where

$$A(x) := [t_i'(x) t_j'(x)]_{ij}$$

is a $D \times D$ symmetric matrix, and

$$b(x) = [t_1''(x) \ldots t_D''(x)]^\top$$

is a D-dimensional column vector.

Theorem 2. *The best conversion of a density $r(x)$ into a density $p_\theta(x)$ of an exponential family minimizing the right-sided Hyvärinen divergence is*

$$\theta_{\mathrm{SME}}(r) = -\Big(E_r[[t_i'(x) t_j'(x)]_{ij}]\Big)^{-1} \times \Big(E_r[[t_1''(x) \ldots t_D''(x)]]^\top\Big).$$

- Solution instantiated for polynomial exponential families:
For polynomial exponential families of order D, we have $t_i'(x) = i x^{i-1}$ and $t_i''(x) = i(i-1)x^{i-2}$, and therefore, we have

$$A_D = E_r[A(x)] = \left[ij\, \mu_{i+j-2}(r)\right]_{ij},$$

and

$$b_D = E_s[b(x)] = \left[j(j-1)\, \mu_{j-2}(r)\right]_j,$$

where $\mu_l(r) := E_r[X^l]$ denotes the l-th raw moment of distribution $X \sim r(x)$ (with the convention that $m_{-1}(r) = 0$). For a probability density function $r(x)$, we have $\mu_1(r) = 1$.
Thus, the integral-based SME of a density r is:

$$\theta_{\mathrm{SME}}(r) = -\left(\left[ij\mu_{i+j-2}(r)\right]_{ij}\right)^{-1} \times \left[j(j-1)\mu_{j-2}(r)\right]_j. \tag{18}$$

For example, matrix A_4 is

$$\begin{bmatrix} \mu_0 & 2\mu_1 & 3\mu_2 & 4\mu_3 \\ 2\mu_1 & 4\mu_2 & 6\mu_3 & 8\mu_4 \\ 3\mu_2 & 6\mu_3 & 9\mu_4 & 12\mu_5 \\ 4\mu_3 & 8\mu_4 & 12\mu_5 & 16\mu_6 \end{bmatrix}.$$

- Faster PEF solutions using Hankel matrices:
The method of Cobb et al. [28] (1983) anticipated the Score Matching method of Hyvärinen (2005). It can be derived from Stein's lemma for exponential families [50]. The integral-based Score Matching method is consistent, i.e., if $r = p_\theta$, then $\bar{\theta}_{\mathrm{SME}} = \theta$: The probabilistic proof for $r(x) = p_e(x)$ is reported as Theorem 2 of [28]. The integral-based proof is based on the property that arbitrary order partial mixed derivatives can be obtained from higher-order partial derivatives with respect to θ_1 [29]:

$$\partial_1^{i_1} \ldots \partial_D^{i_D} F(\theta) = \partial_1^{\sum_{j=1}^D j i_j} F(\theta),$$

where $\partial_i := \frac{\partial}{\partial\theta_i}$.
The complexity of the direct SME method is $O(D^3)$ as it requires the inverse of the $D \times D$-dimensional matrix A_D.
We show how to lower this complexity by reporting an equivalent method (originally presented in [28]) that relies on recurrence relationships between the moments of $p_\theta(x)$ for PEDs. Recall that $\mu_l(r)$ denotes the l-th raw moment $E_r[x^l]$.
Let $A' = [a'_{i+j-2}]_{ij}$ denote the $D \times D$ symmetric matrix with $a'_{i+j-2}(r) = \mu_{i+j-2}(r)$ (with $a'_0(r) = \mu_0(r) = 1$), and $b' = [b_i]_i$ the D-dimensional vector with $b'_i(r) = (i+1)\mu_i(r)$. We solve the system $A'\beta = b'$ to obtain $\beta = A'^{-1}b'$. We then obtain the natural parameter $\bar{\theta}_{\mathrm{SME}}$ from the vector β as

$$\bar{\theta}_{\mathrm{SME}} = \begin{bmatrix} -\frac{\beta_1}{2} \\ \vdots \\ -\frac{\beta_i}{i+1} \\ \vdots \\ -\frac{\beta_D}{D+1} \end{bmatrix}. \quad (19)$$

Now, if we inspect matrix $A'_D = [\mu_{i+j-2}(r)]$, we find that matrix A'_D is a Hankel matrix: A Hankel matrix has constant anti-diagonals and can be inverted in quadratic-time [51,52] instead of cubic time for a general $D \times D$ matrix. (The inverse of a Hankel matrix is a Bezoutian matrix [53].) Moreover, a Hankel matrix can be stored using linear memory (store $2D - 1$ coefficients) instead of quadratic memory of regular matrices.
For example, matrix A'_4 is:

$$A'_4 = \begin{bmatrix} \mu_0 & \mu_1 & \mu_2 & \mu_3 \\ \mu_1 & \mu_2 & \mu_3 & \mu_4 \\ \mu_2 & \mu_3 & \mu_4 & \mu_5 \\ \mu_3 & \mu_4 & \mu_5 & \mu_6 \end{bmatrix},$$

and requires only $6 = 2 \times 4 - 2$ coefficients to be stored instead of $4 \times 4 = 16$. The order-d moment matrix is

$$A'_d := [\mu_{i+j-2}]_{ij} = \begin{bmatrix} \mu_0 & \mu_1 & \ldots & \mu_d \\ \mu_1 & \mu_2 & \ldots & \vdots \\ \vdots & & \ddots & \vdots \\ \mu_d & \ldots & \ldots & \mu_{2d} \end{bmatrix},$$

is a Hankel matrix stored using $2d+1$ coefficients:

$$A'_d =: \mathrm{Hankel}(\mu_0, \mu_1, \ldots, \mu_{2d}).$$

In statistics, those matrices A'_d are called moment matrices and well-studied [54–56]. The variance $\mathrm{Var}[X]$ of a random variable X can be expressed as the determinant of the order-2 moment matrix:

$$\mathrm{Var}[X] = E[(X-\mu)^2] = E[X^2] - E[X]^2 = \mu_2 - \mu_1^2 = \det\left(\begin{bmatrix} 1 & \mu_1 \\ \mu_1 & \mu_2 \end{bmatrix}\right) \geq 0.$$

This observation yields a generalization of the notion of variance to $d+1$ random variables: $X_1, \ldots, X_{d+1} \sim_{iid} F_X \Rightarrow E\left[\prod_{j>i}(X_i - X_j)^2\right] = (d+1)!\det(M_d) \geq 0$. The variance can be expressed as $E[\frac{1}{2}(X_1 - X_2)^2]$ for $X_1, X_2 \sim_{iid} F_X$. See [57] (Chapter 5) for a detailed description related to U-statistics.

For GMMs r, the raw moments $\mu_l(r)$ to build matrix A_D can be calculated in closed-form, as explained in Section 2.4.

Theorem 3 (Score matching GMM conversion)**.** *The Score Matching conversion of a GMM $m(x)$ into a polynomial exponential density $p_{\theta_{\mathrm{SME}}(m)}(x)$ of order D is obtained as*

$$\theta_{\mathrm{SME}}(m) = -\left(\left[ij\, m_{i+j-2}\right]_{ij}\right)^{-1} \times \left[j(j-1)\, m_{j-2}\right]_j,$$

where $m_i = E_m[x^i]$ denote the ith non-central moment of the GMM $m(x)$.

2.3. Converting Numerically Moment Parameters from/to Natural Parameters

Recall that our fast heuristic approximates the Jeffreys divergence by

$$\tilde{D}_J[m, m'] := (\bar{\theta}_{\mathrm{SME}}(m') - \bar{\theta}_{\mathrm{SME}}(m))^\top (\bar{\eta}_{\mathrm{MLE}}(m') - \bar{\eta}_{\mathrm{MLE}}(m)).$$

Because F and ∇F^* are not available in closed form (except for the case $D = 2$ of the normal family), we cannot obtain θ from a given η (using $\theta = \nabla F^*(\eta)$) nor η from a given θ (using $\eta = \nabla F(\theta)$).

However, provided that we can approximate numerically $\tilde{\eta} \simeq \nabla F(\theta)$ and $\tilde{\theta} \simeq \nabla F^*(\eta)$, we also consider these two approximations for the Jeffreys divergence:

$$\tilde{\Delta}_J^{\mathrm{MLE}}[m_1, m_2] := (\tilde{\theta}_2^{\mathrm{MLE}} - \tilde{\theta}_1^{\mathrm{MLE}})^\top (\bar{\eta}_2^{\mathrm{MLE}} - \bar{\eta}_1^{\mathrm{MLE}}),$$

and

$$\tilde{\Delta}_J^{\mathrm{SME}}[m_1, m_2] = (\bar{\theta}_2^{\mathrm{SME}} - \bar{\theta}_1^{\mathrm{SME}})^\top (\tilde{\eta}_2^{\mathrm{SME}} - \tilde{\eta}_1^{\mathrm{SME}}).$$

We show how to numerically estimate $\tilde{\theta}^{\mathrm{MLE}} \simeq \nabla F(\bar{\eta}^{\mathrm{MLE}})$ from $\bar{\eta}^{\mathrm{MLE}}$ in Section 2.3.1. Next, in Section 2.3.2, we show how to stochastically estimate $\tilde{\eta}^{\mathrm{SME}} \simeq \nabla F^*(\bar{\theta}^{\mathrm{SME}})$.

2.3.1. Converting Moment Parameters to Natural Parameters Using Maximum Entropy

Let us report the iterative approximation technique of [58] (which extended the method described in [35]) based on solving a maximum entropy problem (MaxEnt problem). This method will be useful when comparing our fast heuristic $\tilde{D}_J[m,m']$ with the approximations $\tilde{\Delta}_J^{\mathrm{MLE}}[m,m']$ and $\tilde{\Delta}_J^{\mathrm{SME}}[m,m']$.

The density p_θ of any exponential family can be characterized as a maximum entropy distribution given the D moment constraints $E_{p_\theta}[t_i(x)] = \eta_i$: Namely, $\max_p h(p)$ subject to the $D+1$ moment constraints $\int t_i(x)p(x)\mathrm{d}x = \eta_i$ for $i \in \{0,\ldots,D\}$, where we added by convention $\eta_0 = 1$ and $t_0(x) = 1$ (so that $\int p(x)\mathrm{d}x = 1$). The solution of this MaxEnt problem [58] is $p(x) = p_\lambda$, where λ are the $D+1$ Lagrangian parameters. Here, we adopt the following canonical parameterization of the densities of an exponential family:

$$p_\lambda(x) := \exp\left(-\sum_{i=0}^{D} \lambda_i t_i(x)\right).$$

That is, $F(\lambda) = \lambda_0$ and $\lambda_i = -\theta_i$ for $i \in \{1,\ldots,D\}$. Parameter λ is a kind of augmented natural parameter that includes the log-normalizer in its first coefficient.

Let $K_i(\lambda) := E_{p_\theta}[t_i(x)] = \eta_i$ denote the set of $D+1$ non-linear equations for $i \in \{0,\ldots,D\}$. The Iterative Linear System Method [58] (ILSM) converts p^η to p_θ iteratively. We initialize $\lambda^{(0)}$ to $\bar{\theta}_{\mathrm{SME}}$ (and calculate numerically $\lambda_0^{(0)} = F(\bar{\theta}_{\mathrm{SME}})$).

At iteration t with current estimate $\lambda^{(t)}$, we use the following first-order Taylor approximation:

$$K_i(\lambda) \approx K_i(\lambda^{(t)}) + (\lambda - \lambda^{(t)})\nabla K_i(\lambda^{(t)}).$$

Let $H(\lambda)$ denote the $(D+1) \times (D+1)$ matrix:

$$H(\lambda) := \left[\frac{\partial K_i(\lambda)}{\partial \theta_j}\right]_{ij}.$$

We have

$$H_{ij}(\lambda) = H_{ji}(\lambda) = -E_{p_\theta}[t_i(x)t_j(x)].$$

We update as follows:

$$\lambda^{(t+1)} = \lambda^{(t)} + H^{-1}(\lambda^{(t)}) \begin{bmatrix} \eta_0 - K_0(\lambda^{(t)}) \\ \vdots \\ \eta_D - K_D(\lambda^{(t)}) \end{bmatrix}. \tag{20}$$

For a PEF of order D, we have

$$H_{ij}(\lambda) = -E_{p_\theta}[x^{i+j-2}] = -\mu_{i+j-2}(p_\theta).$$

This yields a moment matrix H_λ (Hankel matrix), which can be inverted in quadratic time [52]. In our setting, the moment matrix is invertible because $|H| > 0$, see [59].

Let $\tilde{\lambda}_T(\eta)$ denote $\theta^{(T)}$ after T iterations (retrieved from $\lambda^{(T)}$) and the corresponding natural parameter of the PED. We have the following approximation of the JD:

$$D_J[m,m'] \approx (\tilde{\theta}_T(\eta') - \tilde{\theta}_T(\eta))^\top(\eta' - \eta).$$

The method is costly because we need to numerically calculate $\mu_{i+j-2}(p_\theta)$ and the K_i's (e.g., univariate Simpson integrator). Another potential method consists of estimating these expectations using acceptance-rejection sampling [60,61]. We may also consider the holonomic gradient descent [29]. Thus, the conversion $\eta \to \theta$ method is costly. Our heuristic $\tilde{\Delta}_J$ bypasses this costly moment-to-natural parameter conversion by converting each mixture m to a pair $(p_{\theta_{\mathrm{SME}}}, p_{\eta_{\mathrm{MLE}}})$ of PEDs parameterized in the natural and moment parameters (i.e., loosely speaking, we untangle these dual parameterizations).

2.3.2. Converting Natural Parameters to Moment Parameters

Given a PED $p_\theta(x)$, we have to find its corresponding moment parameter η (i.e., $p_\theta = p^\eta$). Since $\eta = E_{p_\theta}[t(x)]$, we sample s i.i.d. variates $x_1, \ldots, x_s$ from p_θ using acceptance-rejection sampling [60,61] or any other Markov chain Monte Carlo technique [62] and estimate $\hat{\eta}$ as:

$$\hat{\eta} = \frac{1}{s}\sum_{i=1}^{s} t(x_i).$$

2.4. Raw Non-Central Moments of Normal Distributions and GMMs

In order to implement the MLE or SME Gaussian mixture conversion procedures, we need to calculate the raw moments of a Gaussian mixture model. The l-th moment raw moment $E[Z^l]$ of a standard normal distribution $Z \sim N(0,1)$ is 0 when l is odd (since the normal standard density is an even function) and $(l-1)!! = 2^{-\frac{l}{2}}\frac{l!}{(l/2)!}$ when l is even, where $n!! = \sqrt{\frac{2^{n+1}}{\pi}}\Gamma(\frac{n}{2}+1) = \prod_{k=0}^{\lceil \frac{n}{2}\rceil -1}(n-2k)$ is the double factorial (with $(-1)!! = 1$ by convention). Using the binomial theorem, we deduce that a normal distribution $X = \mu + \sigma Z$ has finite moments:

$$\mu_l(p_{\mu,\sigma}) = E_{p_{\mu,\sigma}}[X^l] = E[(\mu+\sigma Z)^l] = E[(\mu+\sigma Z)^l] = \sum_{i=0}^{l}\binom{l}{i}\mu^{l-i}\sigma^i E[Z^i].$$

That is, we have

$$\mu_l(p_{\mu,\sigma}) = \sum_{i=0}^{\lfloor \frac{l}{2}\rfloor}\binom{l}{i}(2i-1)!!\,\mu^{l-2i}\sigma^{2i}, \tag{21}$$

where $n!!$ denotes the double factorial:

$$n!! = \prod_{k=0}^{\lceil \frac{n}{2}\rceil -1}(n-2k) = \begin{cases} \prod_{k=1}^{\frac{n}{2}}(2k) & n \text{ is even,} \\ \prod_{k=1}^{\frac{n+1}{2}}(2k-1) & n \text{ is odd.} \end{cases}$$

By the linearity of the expectation $E[\cdot]$, we deduce the l-th raw moment of a GMM $m(x) = \sum_{i=1}^{k} w_i p_{\mu_i,\sigma_i}(x)$:

$$\mu_l(m) = \sum_{i=1}^{k} w_i \mu_l(p_{\mu_I,\sigma_i}).$$

Notice that by using [63], we can extend this formula to truncated normals and GMMs. Thus, computing the first $O(D)$ raw moments of a GMM with k components can be done in $O(kD^2)$ using the Pascal triangle method for computing the binomial coefficients. See also [64].

3. Goodness-of-Fit between GMMs and PEDs: Higher Order Hyvärinen Divergences

Once we have converted a GMM $m(x)$ into an unnormalized PED $q_{\theta_m}(x) = \tilde{p}_{\theta_m}(x)$, we would like to evaluate the quality of the conversion, i.e., $D[m(x) : q_{\theta_m}(x)]$, using a statistical divergence $D[\cdot : \cdot]$. This divergence shall allow us to perform model selection by choosing the order D of the PEF so that $D[m(x) : p_\theta(x)] \leq \epsilon$ for $\theta \in \mathbb{R}^D$, where $\epsilon > 0$ is a prescribed threshold. Since PEDs have computationally intractable normalization constants, we consider a right-sided projective divergence [42] $D[p : q]$ that satisfies $D[p : \lambda q] = D[p : q] = D[p : \tilde{q}]$ for any $\lambda > 0$. For example, we may consider the γ-divergence [65] that is a two-sided projective divergence: $D_\gamma[\lambda p : \lambda' q] = D[p : q] = D[\tilde{p} : \tilde{q}]$ for any $\lambda, \lambda' > 0$ and converge to the KLD when $\gamma \to 0$. However, the γ-divergence between a mixture model and an unnormalized PEF does not yield a closed-form formula. Moreover, the γ-divergence between two unnormalized PEDs is expressed using the log-normalizer function $F(\cdot)$ that is computationally intractable [66].

In order to a get a closed-form formula for a divergence between a mixture model and an unnormalized PED, we consider the order-α (for $\alpha > 0$) Hyvärinen divergence [42] as follows:

$$D_{H,\alpha}[p:q] := \int p(x)^\alpha \left(\nabla_x \log p(x) - \nabla_x \log q(x)\right)^2 \mathrm{d}x, \quad \alpha > 0. \tag{22}$$

The Hyvärinen divergence [42] (order-1 Hyvärinen divergence) has also been called the Fisher divergence [27,67–69] or relative Fisher information [48]. Notice that when $\alpha = 1$, $D_{H,1}[p:q] = D_H[p:q]$, the ordinary Hyvärinen divergence [27].

The Hyvärinen divergences $D_{H,\alpha}$ is a right-sided projective divergence, meaning that the divergence satisfies $D_{H,\alpha}[p:q] = D_{H,\alpha}[p:\lambda q]$ for any $\lambda > 0$. That is, we have $D_{H,\alpha}[p:q] = D_{H,\alpha}[p:\tilde{q}]$. Thus, we have $D_{H,\alpha}[m:p_\theta] = D_{H,\alpha}[m:q_\theta]$ for an unnormalized PED $q_\theta = \tilde{p}_\theta$. For statistical estimation, it is enough to have a sided projective divergence since we need to evaluate the goodness of fit between the (normalized) empirical distribution p_e and the (unnormalized) parameteric density.

For univariate distributions, $\nabla_x \log p(x) = \frac{p'(x)}{p(x)}$, and $\frac{p'(x)}{p(x)} = \frac{\tilde{p}'(x)}{\tilde{p}(x)}$, where $\tilde{p}(x)$ is the unnormalized model.

Let $P_\theta(x) := \sum_{i=1}^D \theta_i x^i$ be a homogeneous polynomial defining the shape of the EPF:

$$p_\theta(x) = \exp(P_\theta(x) - F(\theta)).$$

For PEDs with the homogeneous polynomial $P_\theta(x)$, we have $\frac{p'(x)}{p(x)} = (\log P_\theta(x))' = \sum_{i=1}^D i\theta_i x^{i-1}$.

Theorem 4. *The Hyvärinen divergence $D_{H,2}[m:q_\theta]$ of order 2 between a Gaussian mixture $m(x)$ and a polynomial exponential family density $q_\theta(x)$ is available in closed form.*

Proof. We have $D_{H,2}[m:q] = \int m(x)^2 \left(\frac{m'(x)}{m(x)} - \sum_{i=1}^D i\theta_i x^{i-1}\right)^2 \mathrm{d}x$ with

$$m'(x) = -\sum_{i=1}^k w_i \frac{x-\mu_i}{\sigma_i^2} p(x_i;\mu_i,\sigma_i),$$

denoting the derivative of the Gaussian mixture density $m(x)$. It follows that:

$$D_{H,2}[m:q] = \int m'(x)\mathrm{d}x - 2\sum_{i=1}^D i\theta_i \int x^{i-1} m'(x)m(x)\mathrm{d}x + \sum_{i,j=1}^D ij\theta_i\theta_j \int x^{i+j-2} m(x)^2 \mathrm{d}x,$$

where

$$\int x^i m'(x)m(x)\mathrm{d}x = -\sum w_a w_b \int \frac{x-\mu_a}{\sigma_a^2} x^i p(x;\mu_a,\sigma_a)p(x;\mu_b,\sigma_b)\mathrm{d}x.$$

Therefore, we have

$$D_{H,2}[m:q] = \int m'(x)\mathrm{d}x - 2\sum_{i=1}^D i\theta_i \int x^{i-1} m'(x)m(x)\mathrm{d}x + \sum_{i,j=1}^D ij\theta_i\theta_j \int x^{i+j-2} m(x)^2 \mathrm{d}x$$

with $m'(x) = -\sum w_a \frac{x-\mu_a}{\sigma_a^2} p(x;\mu_a,\sigma_a)$.

Since $p_a(x)p_b(x) = \kappa_{a,b} p(x;\mu_{ab},\sigma_{ab})$, with

$$\begin{aligned}
\mu_{ab} &= \sigma_a^2\sigma_b^2(\sigma_b^2\mu_a + \sigma_a^2\mu_b), \\
\sigma_{ab} &= \frac{\sigma_a\sigma_b}{\sqrt{\sigma_a^2+\sigma_b^2}}, \\
\kappa_{a,b} &= \exp(F(\mu_{ab},\sigma_{ab}) - F(\mu_a,\sigma_a) - F(\mu_b,\sigma_b)),
\end{aligned}$$

and

$$F(\mu, \sigma) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2),$$

the log-normalizer of the Gaussian exponential family [42].

Therefore, we obtain

$$\int p_a(x) p_b(x) x^l \mathrm{d}x = \kappa_{a,b} m_l(\mu_{ab}, \sigma_{ab}).$$

Thus, the Hyvärinen divergence $D_{H,2}$ of order 2 between a GMM and a PED is available in closed-form. □

For example, when $k = 1$ (i.e., mixture m is a single Gaussian p_{μ_1,σ_1}) and p_θ is a normal distribution (i.e., PED with $D = 2$, $q_\theta = p_{\mu_2,\sigma_2}$), we obtain the following formula for the order-2 Hyvärinen divergence:

$$D_{H,2}[p_{\mu_1,\sigma_1} : p_{\mu_2,\sigma_2}] = \frac{(\sigma_1^2 - \sigma_2^2)^2 + 2(\mu_2 - \mu_1)^2\sigma_1^2}{8\sqrt{\pi}\sigma_1^3\sigma_2^4}.$$

4. Experiments: Jeffreys Divergence between Mixtures

In this section, we evaluate our heuristic to approximate the Jeffreys divergence between two mixtures m and m':

$$\tilde{D}_J[m, m'] := (\bar{\theta}_{\mathrm{SME}}(m') - \bar{\theta}_{\mathrm{SME}}(m))^\top (\bar{\eta}_{\mathrm{MLE}}(m') - \bar{\eta}_{\mathrm{MLE}}(m)).$$

Recall that stochastically estimating the JD between k-GMMs with Monte Carlo sampling using s samples (i.e., $\hat{D}_{J,s}[m : m']$) requires $\tilde{O}(ks)$ and is not deterministic. That is, different MC runs yield fluctuating values that may be fairly different. In comparison, approximating D_J by $\tilde{D}_J$ using Δ_J by converting mixtures to D-order PEDs require $O(kD^2)$ time to compute the raw moments and $O(D^2)$ time to invert a Hankel moment matrix. Thus, by choosing $D = 2k$, we obtain a deterministic $O(k^3)$ algorithm that is faster than the MC sampling when $k^2 \ll s$. Since there are, at most, k modes for a k-GMM, we choose order $D = 2k$ for the PEDs.

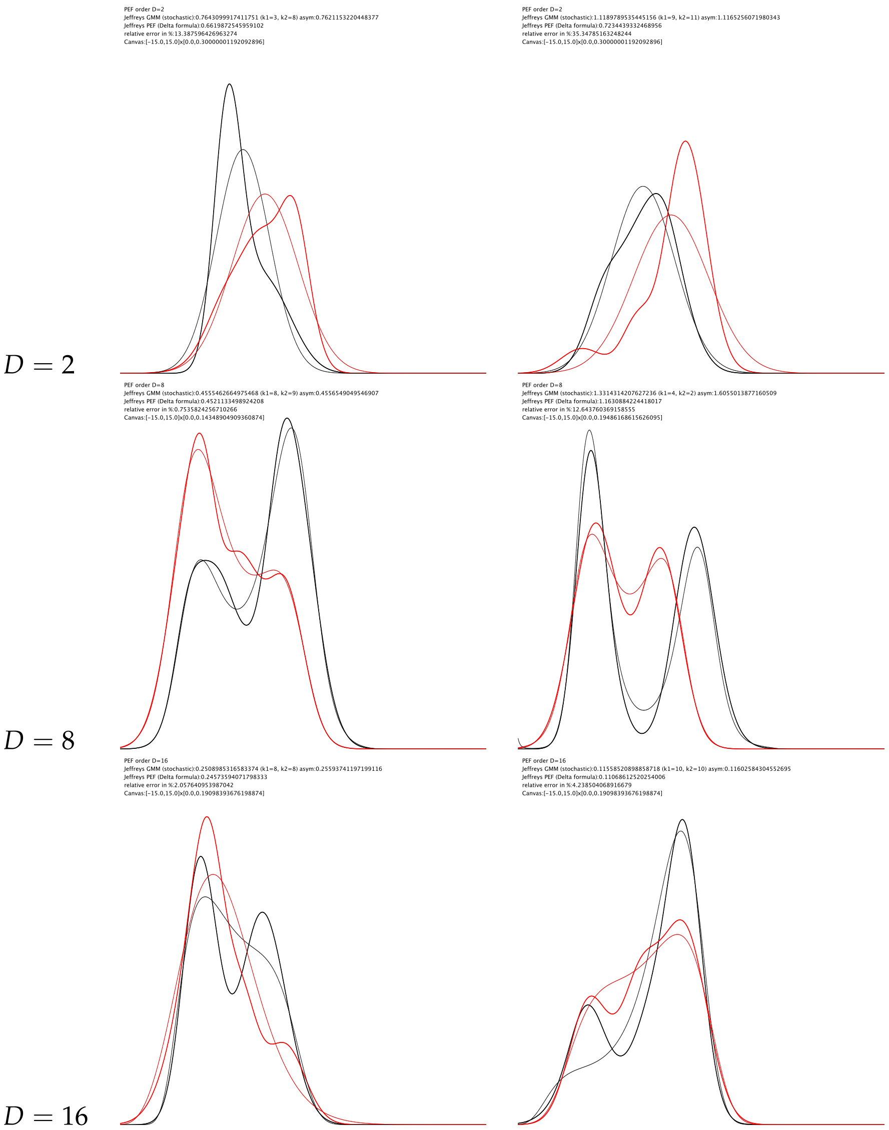
To obtain quantitative results on the performance of our heuristic $\tilde{D}_J$, we build random GMMs with k components as follows: $m(x) = \sum_{i=1}^k w_i p_{\mu_i,\sigma_i}(x)$, where $w_i \sim U_i$, $\mu_i \sim -10 + 10U_1'$ and $\sigma_i \sim 1 + U_2'$, where the U_i's, and U_1' and U_2' are independent uniform distributions on $[0, 1)$. The mixture weights are then normalized to sum up to one. For each value of k, we make 1000 trial experiments to gather statistics and use $s = 10^5$ for evaluating the Jeffreys divergence $\hat{D}_J$ by Monte Carlo samplings. We denote by error $:= \frac{|\hat{D}_J - \Delta_J|}{\hat{D}_J}$ the error of an experiment. Table 1 presents the results of the experiments for $D = 2k$: The table displays the average error, the maximum error (minimum error is very close to zero, of order 10^{-5}), and the speed-up obtained by our heuristic Δ_J. Those experiments were carried out on a Dell Inspiron 7472 laptop (equipped with an Intel(R) Core(TM) i5-8250U CPU at 1.60 GHz).

Table 1. Comparison of $\tilde{\Delta}_J(m_1, m_2)$ with $\hat{D}_J(m_1, m_2)$ for random GMMs.

k	D	**Average Error**	**Maximum Error**	**Speed-Up**
2	4	0.1180799978221536	0.9491425404132259	2008.2323536011806
3	6	0.12533811294546526	1.9420608151988419	1010.4917042114389
4	8	0.10198448868508087	5.290871019594698	474.5135294829539
5	10	0.06336388579897352	3.8096955246161848	246.38780782640987
6	12	0.07145257192133717	1.0125283726458822	141.39097909641052
7	14	0.10538875853178625	0.8661463142793943	88.62985036546912
8	16	0.4150905507007969	0.4150905507007969	58.72277575395611

Notice that the quality of the approximations of $\tilde{D}_J$ depend on the number of modes of the GMMs. However, calculating the number of modes is difficult [43,70], even for simple cases [71,72].

Figure 4 displays several experiments of converting mixtures to pairs of PEDs to obtain approximations of the Jeffreys divergence.



```
PEF order D=2
Jeffreys GMM (stochastic):0.7643099917411751 (k1=3, k2=8) asym:0.7621153220448377
Jeffreys PEF (Delta formula):0.6619872545959102
relative error in %:13.387596426963274
Canvas:[-15.0,15.0]x[0.0,0.30000001192092896]
```

```
PEF order D=2
Jeffreys GMM (stochastic):1.1189789535445156 (k1=9, k2=11) asym:1.1165256071980343
Jeffreys PEF (Delta formula):0.7234439332468956
relative error in %:35.34785163248244
Canvas:[-15.0,15.0]x[0.0,0.30000001192092896]
```

```
PEF order D=8
Jeffreys GMM (stochastic):0.4555462664975468 (k1=8, k2=9) asym:0.4556549049546907
Jeffreys PEF (Delta formula):0.4521133498924208
relative error in %:0.7535824256710266
Canvas:[-15.0,15.0]x[0.0,0.14348904909360874]
```

```
PEF order D=8
Jeffreys GMM (stochastic):1.3314314207627236 (k1=4, k2=2) asym:1.6055013877160509
Jeffreys PEF (Delta formula):1.1630884224418017
relative error in %:12.643760369158555
Canvas:[-15.0,15.0]x[0.0,0.19486168615626095]
```

```
PEF order D=16
Jeffreys GMM (stochastic):0.2508985316583374 (k1=8, k2=8) asym:0.25593741197199116
Jeffreys PEF (Delta formula):0.24573594071798333
relative error in %:2.057640953987042
Canvas:[-15.0,15.0]x[0.0,0.19098393676198874]
```

```
PEF order D=16
Jeffreys GMM (stochastic):0.11558520898858718 (k1=10, k2=10) asym:0.11602584304552695
Jeffreys PEF (Delta formula):0.11068612520254006
relative error in %:4.238504068916679
Canvas:[-15.0,15.0]x[0.0,0.19098393676198874]
```

Figure 4. Experiments of approximating the Jeffreys divergence between two mixtures by considering pairs of PEDs. Notice that only the PEDs estimated using the Score Matching in the natural parameter space are displayed.

Figure 5 illustrates the use of the order-2 Hyvärinen divergence $D_{H,2}$ to perform model selection for choosing the order of a PED.

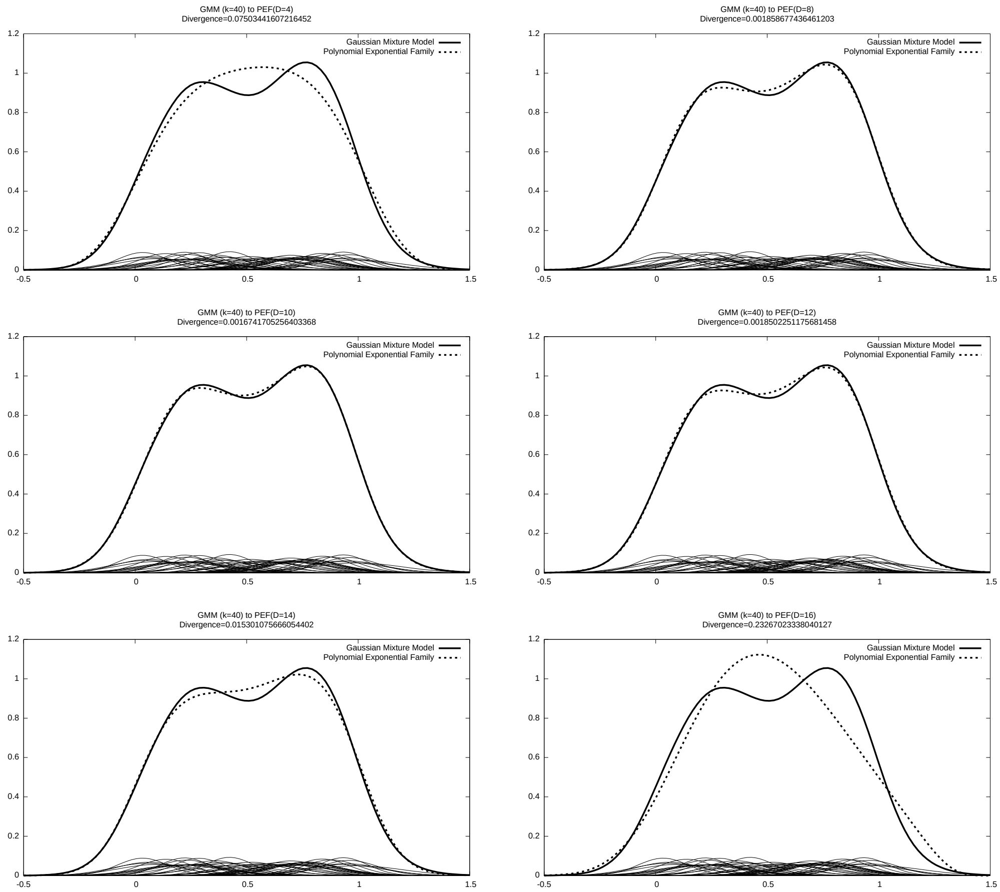


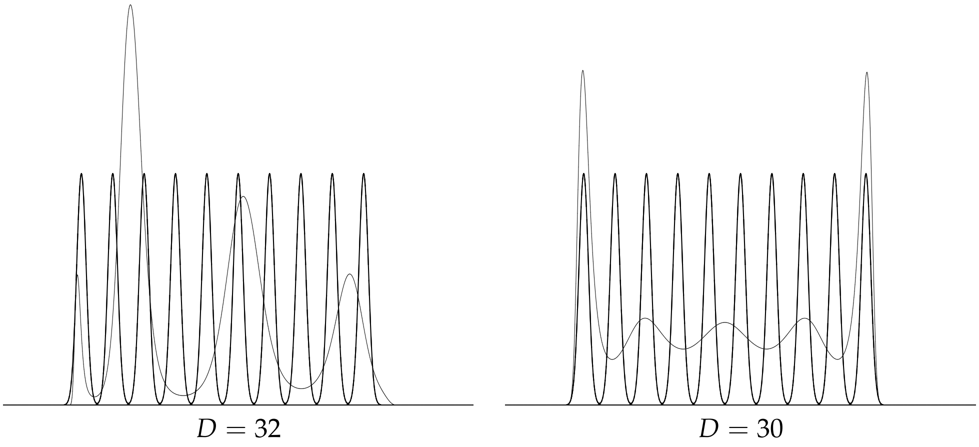
Figure 5. Selecting the PED order D my evaluating the best divergence order-2 Hyvärinen divergence (for $D \in \{4, 8, 10, 12, 14, 16\}$) values. Here, the order $D = 10$ (boxed) yields the lowest order-2 Hyvärinen divergence: The GMM is close to the PED.

Finally, Figure 6 displays some limitations of the GMM to PED conversion when the GMMs have many modes. In that case, running the conversion $\bar{\eta}_{\mathrm{MLE}}$ to obtain $\tilde{\theta}_T(\bar{\eta}_{\mathrm{MLE}})$ and estimate the Jeffreys divergence by

$$\tilde{\Delta}_J^{\mathrm{MLE}}[m_1, m_2] = (\tilde{\theta}_2^{\mathrm{MLE}} - \tilde{\theta}_1^{\mathrm{MLE}})^\top (\bar{\eta}_2^{\mathrm{MLE}} - \bar{\eta}_1^{\mathrm{MLE}}),$$

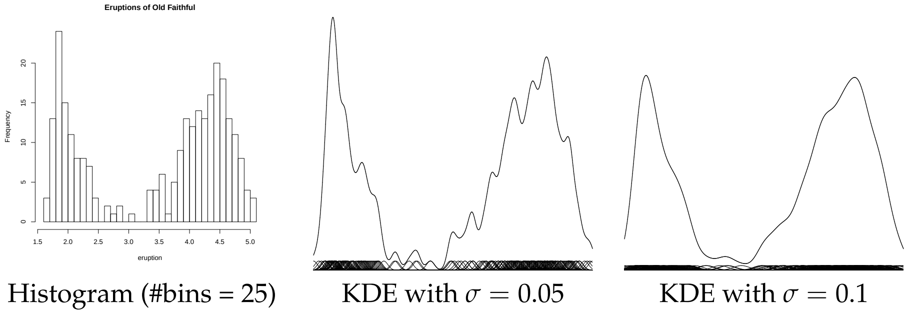
improves the results but requires more computation.

Next, we consider learning a PED by converting a GMM derived itself from a Kernel Density Estimator (KDE). We use the duration of the eruption for the Old Faithful geyser in Yellowstone National Park (Wyoming, USA): The dataset consists of 272 observations (https://www.stat.cmu.edu/~larry/all-of-statistics/=data/faithful.dat) (access date: 25 October 2021) and is included in the R language package 'stats'. Figure 7 displays the GMMs obtained from the KDEs of the Old Faithful geyser dataset when choosing for each component $\sigma = 0.05$ (left) and $\sigma = 0.1$. Observe that the data are bimodal once the spurious modes (i.e., small bumps) are removed, as studied in [32]. Barron and Sheu [32] modeled that dataset using a bimodal PED of order $D = 4$, i.e., a quartic distribution. We model it with a PED of order $D = 10$ using the integral-based Score Matching method. Figure 8 displays the unnormalized bimodal density q_1 (i.e., $\tilde{p}_1$) that we obtained using the integral-based Score Matching method (with $\mathcal{X} = (0, 1)$).



$D = 32$ $D = 30$

Figure 6. Some limitation examples of the conversion of GMMs (black) to PEDs (grey) using the integral-based Score Matching estimator: Case of GMMs with many modes.



Histogram (#bins = 25) KDE with $\sigma = 0.05$ KDE with $\sigma = 0.1$

Figure 7. Modeling the Old Faithful geyser by a KDE (GMM with $k = 272$ components, uniform weights $w_i = \frac{1}{272}$): Histogram (#bins = 25) (**left**), KDE with $\sigma = 0.05$ (**middle**), and KDE with $\sigma = 0.1$ with less spurious bumps (**right**).

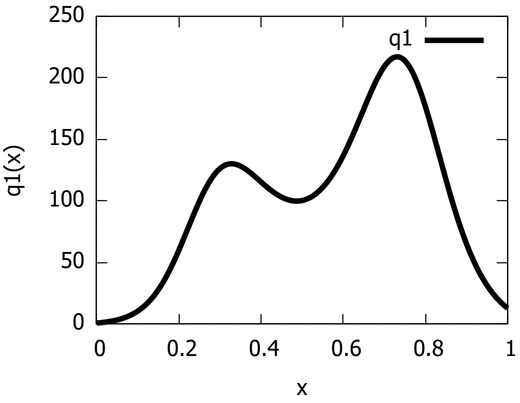


Figure 8. Modeling the Old Faithful geyser by an exponential-polynomial distribution of order $D = 10$.

5. Conclusions and Perspectives

Many applications [7,73–75] require computing the Jeffreys divergence (the arithmetic symmetrization of the Kullback–Leibler divergence) between Gaussian mixture models. Since the Jeffreys divergence between GMMs is provably not available in closed-form [13], one often ends up implementing a costly Monte Carlo stochastic approximation of the Jeffreys divergence. In this paper, we first noticed the simple expression of the Jeffreys divergence between densities p_θ and $p_{\theta'}$ of an exponential family using their dual natural and moment parameterizations [22] $p_\theta = p^\eta$ and $p_{\theta'} = p^{\eta'}$:

$$D_J[p_\theta, p_{\theta'}] = (\theta' - \theta)^\top (\eta' - \eta),$$

where $\eta = \nabla F(\theta)$ and $\eta' = \nabla F(\theta')$ for the cumulant function $F(\theta)$ of the exponential family. This led us to propose a simple and fast heuristic to approximate the Jeffreys divergence between Gaussian mixture models: First, convert a mixture m to a pair $(p_{\bar{\theta}^{\mathrm{SME}}}, p^{\bar{\eta}^{\mathrm{MLE}}})$ of dually parameterized polynomial exponential densities using extensions of the Maximum Likelihood and Score Matching Estimators (Theorems 1 and 3), and then approximate the JD deterministically by

$$D_J[m_1, m_2] \simeq \tilde{D}_J[m_1, m_2] = (\tilde{\theta}_2^{\mathrm{MLE}} - \tilde{\theta}_1^{\mathrm{MLE}})^\top (\bar{\eta}_2^{\mathrm{MLE}} - \bar{\eta}_1^{\mathrm{MLE}}).$$

The order of the polynomial exponential family may be either prescribed or selected using the order-2 Hyvärinen divergence, which evaluates in closed form the dissimilarity between a GMM and a density of an exponential-polynomial family (Theorem 4). We experimentally demonstrated that the Jeffreys divergence between GMMs can be reasonably well approximated by $\tilde{D}_J$ for mixtures with a small number of modes, and we obtained an overall speed-up of several order of magnitudes compared to the Monte Carlo sampling method. We also propose another deterministic heuristic to estimate D_J as

$$\tilde{D}_J^{\mathrm{MLE}}[m_1 : m_2] = (\tilde{\theta}_2^{\mathrm{MLE}} - \tilde{\theta}_1^{\mathrm{MLE}})^\top (\bar{\eta}_2^{\mathrm{MLE}} - \bar{\eta}_1^{\mathrm{MLE}}),$$

where $\tilde{\theta}^{\mathrm{MLE}} \approx \nabla F(\bar{\eta}^{\mathrm{MLE}})$ is numerically calculated using an iterative conversion procedure based on maximum entropy [58] (Section 2.3.1). Our technique extends to other univariate mixtures of exponential families (e.g., mixtures of Rayleigh distributions, mixtures of Gamma distributions, or mixtures of Beta distributions, etc). One limitation of our method is that the PED modeling of a GMM may not guarantee obtaining the same number of modes as the GMM even when we increase the order D of the exponential-polynomial densities. This case is illustrated in Figure 9 (right).

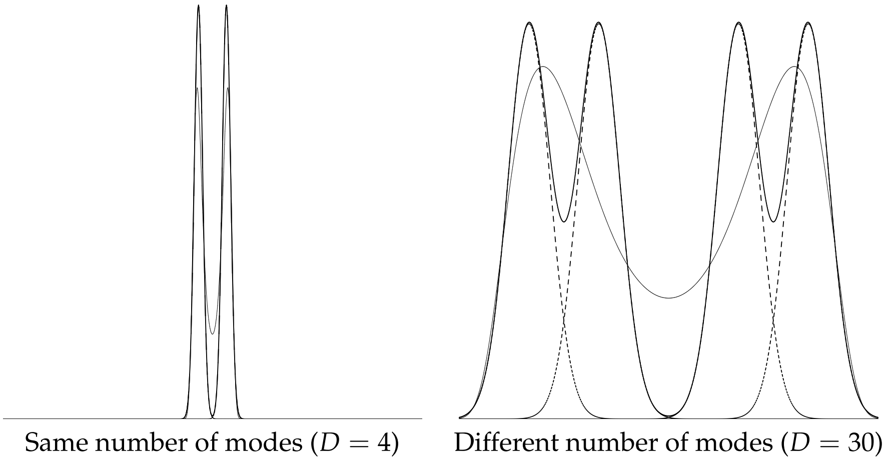


Figure 9. GMM modes versus PED modes: (**left**) same number and locations of modes for the GMM and the PED; (**right**) 4 modes for the GMM but only 2 modes for the PED.

Although PEDs are well-suited to calculate Jeffreys divergence compared to GMMs, we point out that GMMs are better suited for sampling, while PEDs require Monte Carlo methods (e.g., adaptive rejection sampling or MCMC methods [62]). Furthermore, we can estimate the Kullback–Leibler divergence between two PEDs using rejection sampling (or other McMC methods [62]) or by using the γ-divergence [76] with γ close to zero [66] (e.g., $\gamma = 0.001$). The web page of the project is https://franknielsen.github.io/JeffreysDivergenceGMMPEF/index.html (accessed on 25 October 2021).

This work opens up several perspectives for future research: For example, we may consider bivariate polynomial-exponential densities for modeling bivariate Gaussian mixture models [29], or we may consider truncating the GMMs in order to avoid tail phenomena when converting GMMs to PEDs [77,78].

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **1946**, *186*, 453–461.
2. McLachlan, G.J.; Basford, K.E. *Mixture Models: Inference and Applications to Clustering*; M. Dekker: New York, NY, USA, 1988; Volume 38.
3. Pearson, K. Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. Lond. A* **1894**, *185*, 71–110.
4. Seabra, J.C.; Ciompi, F.; Pujol, O.; Mauri, J.; Radeva, P.; Sanches, J. Rayleigh mixture model for plaque characterization in intravascular ultrasound. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 1314–1324. [CrossRef] [PubMed]
5. Kullback, S. *Information Theory and Statistics*; Courier Corporation: North Chelmsford, MA, USA, 1997.
6. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
7. Vitoratou, S.; Ntzoufras, I. Thermodynamic Bayesian model comparison. *Stat. Comput.* **2017**, *27*, 1165–1180. [CrossRef]
8. Kannappan, P.; Rathie, P. An axiomatic characterization of J-divergence. In *Transactions of the Tenth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*; Springer: Dordrecht, The Netherlands, 1988; pp. 29–36.
9. Burbea, J. J-Divergences and related concepts. *Encycl. Stat. Sci.* **2004**. doi:10.1002/0471667196.ess1304
10. 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. [CrossRef]
11. Veldhuis, R. The centroid of the symmetrical Kullback-Leibler distance. *IEEE Signal Process. Lett.* **2002**, *9*, 96–99. [CrossRef]
12. 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. [CrossRef]
13. Watanabe, S.; Yamazaki, K.; Aoyagi, M. Kullback information of normal mixture is not an analytic function. *IEICE Tech. Rep. Neurocomput.* **2004**, *104*, 41–46.
14. Cui, S.; Datcu, M. Comparison of Kullback-Leibler divergence approximation methods between Gaussian mixture models for satellite image retrieval. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 3719–3722.
15. Cui, S. Comparison of approximation methods to Kullback–Leibler divergence between Gaussian mixture models for satellite image retrieval. *Remote Sens. Lett.* **2016**, *7*, 651–660. [CrossRef]
16. Sreekumar, S.; Zhang, Z.; Goldfeld, Z. Non-asymptotic Performance Guarantees for Neural Estimation of f-Divergences. In Proceedings of the International Conference on Artificial Intelligence and Statistics (PMLR 2021), San Diego, CA, USA, 18–24 July 2021; pp. 3322–3330.
17. Durrieu, J.L.; Thiran, J.P.; Kelly, F. Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian mixture models. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4833–4836.
18. Nielsen, F.; Sun, K. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy* **2016**, *18*, 442. [CrossRef]
19. Jenssen, R.; Principe, J.C.; Erdogmus, D.; Eltoft, T. The Cauchy–Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *J. Frankl. Inst.* **2006**, *343*, 614–629. [CrossRef]
20. Liu, M.; Vemuri, B.C.; Amari, S.i.; Nielsen, F. Shape retrieval using hierarchical total Bregman soft clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2407–2419.
21. Robert, C.; Casella, G. *Monte Carlo Statistical Methods*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
22. Barndorff-Nielsen, O. *Information and Exponential Families: In Statistical Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
23. Azoury, K.S.; Warmuth, M.K. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.* **2001**, *43*, 211–246. [CrossRef]
24. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
25. Bregman, L.M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217. [CrossRef]

26. Nielsen, F.; Nock, R. Sided and symmetrized Bregman centroids. *IEEE Trans. Inf. Theory* **2009**, *55*, 2882–2904. [CrossRef]
27. Hyvärinen, A. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **2005**, *6*, 695–709.
28. Cobb, L.; Koppstein, P.; Chen, N.H. Estimation and moment recursion relations for multimodal distributions of the exponential family. *J. Am. Stat. Assoc.* **1983**, *78*, 124–130. [CrossRef]
29. Hayakawa, J.; Takemura, A. Estimation of exponential-polynomial distribution by holonomic gradient descent. *Commun. Stat.-Theory Methods* **2016**, *45*, 6860–6882. [CrossRef]
30. Nielsen, F.; Nock, R. MaxEnt upper bounds for the differential entropy of univariate continuous distributions. *IEEE Signal Process. Lett.* **2017**, *24*, 402–406. [CrossRef]
31. Matz, A.W. Maximum likelihood parameter estimation for the quartic exponential distribution. *Technometrics* **1978**, *20*, 475–484. [CrossRef]
32. Barron, A.R.; Sheu, C.H. Approximation of density functions by sequences of exponential families. *Ann. Stat.* **1991**, *19*, 1347–1369; Correction in **1991**, *19*, 2284–2284.
33. O'toole, A. A method of determining the constants in the bimodal fourth degree exponential function. *Ann. Math. Stat.* **1933**, *4*, 79–93. [CrossRef]
34. Aroian, L.A. The fourth degree exponential distribution function. *Ann. Math. Stat.* **1948**, *19*, 589–592. [CrossRef]
35. Zellner, A.; Highfield, R.A. Calculation of maximum entropy distributions and approximation of marginal posterior distributions. *J. Econom.* **1988**, *37*, 195–209. [CrossRef]
36. McCullagh, P. Exponential mixtures and quadratic exponential families. *Biometrika* **1994**, *81*, 721–729. [CrossRef]
37. Mead, L.R.; Papanicolaou, N. Maximum entropy in the problem of moments. *J. Math. Phys.* **1984**, *25*, 2404–2417. [CrossRef]
38. Armstrong, J.; Brigo, D. Stochastic filtering via L_2 projection on mixture manifolds with computer algorithms and numerical examples. *arXiv* **2013**, arXiv:1303.6236.
39. Efron, B.; Hastie, T. *Computer Age Statistical Inference*; Cambridge University Press: Cambridge, UK, 2016; Volume 5.
40. Pinsker, M. *Information and Information Stability of Random Variables and Processes (Translated and Annotated by Amiel Feinstein)*; Holden-Day Inc.: San Francisco, CA, USA, 1964.
41. Fedotov, A.A.; Harremoës, P.; Topsoe, F. Refinements of Pinsker's inequality. *IEEE Trans. Inf. Theory* **2003**, *49*, 1491–1498. [CrossRef]
42. Amari, S. *Information Geometry and Its Applications*; Applied Mathematical Sciences; Springer: Berlin/Heidelberg, Germany, 2016.
43. Carreira-Perpinan, M.A. Mode-finding for mixtures of Gaussian distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1318–1323. [CrossRef]
44. Brown, L.D. Fundamentals of statistical exponential families with applications in statistical decision theory. *Lect. Notes-Monogr. Ser.* **1986**, *9*, 1–279.
45. Pelletier, B. Informative barycentres in statistics. *Ann. Inst. Stat. Math.* **2005**, *57*, 767–780. [CrossRef]
46. Améndola, C.; Drton, M.; Sturmfels, B. Maximum likelihood estimates for Gaussian mixtures are transcendental. In Proceedings of the International Conference on Mathematical Aspects of Computer and Information Sciences, Berlin, Germany, 11–13 November 2015; pp. 579–590.
47. Hyvärinen, A. Some extensions of score matching. *Comput. Stat. Data Anal.* **2007**, *51*, 2499–2512. [CrossRef]
48. Otto, F.; Villani, C. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.* **2000**, *173*, 361–400. [CrossRef]
49. Toscani, G. Entropy production and the rate of convergence to equilibrium for the Fokker-Planck equation. *Q. Appl. Math.* **1999**, *57*, 521–541. [CrossRef]
50. Hudson, H.M. A natural identity for exponential families with applications in multiparameter estimation. *Ann. Stat.* **1978**, *6*, 473–484. [CrossRef]
51. Trench, W.F. An algorithm for the inversion of finite Hankel matrices. *J. Soc. Ind. Appl. Math.* **1965**, *13*, 1102–1107. [CrossRef]
52. Heinig, G.; Rost, K. Fast algorithms for Toeplitz and Hankel matrices. *Linear Algebra Its Appl.* **2011**, *435*, 1–59. [CrossRef]
53. Fuhrmann, P.A. Remarks on the inversion of Hankel matrices. *Linear Algebra Its Appl.* **1986**, *81*, 89–104. [CrossRef]
54. Lindsay, B.G. On the determinants of moment matrices. *Ann. Stat.* **1989**, *17*, 711–721. [CrossRef]
55. Lindsay, B.G. Moment matrices: applications in mixtures. *Ann. Stat.* **1989**, *17*, 722–740. [CrossRef]
56. Provost, S.B.; Ha, H.T. On the inversion of certain moment matrices. *Linear Algebra Its Appl.* **2009**, *430*, 2650–2658. [CrossRef]
57. Serfling, R.J. *Approximation Theorems of Mathematical Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 162.
58. Mohammad-Djafari, A. A Matlab program to calculate the maximum entropy distributions. In *Maximum Entropy and Bayesian Methods*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 221–233.
59. Karlin, S. *Total Positivity*; Stanford University Press: Redwood City, CA, USA 1968; Volume 1.
60. von Neumann, J. Various Techniques Used in Connection with Random Digits. In *Monte Carlo Method*; National Bureau of Standards Applied Mathematics Series; Householder, A.S., Forsythe, G.E., Germond, H.H., Eds.; US Government Printing Office: Washington, DC, USA, 1951; Volume 12, Chapter 13, pp. 36–38.
61. Flury, B.D. Acceptance-rejection sampling made easy. *SIAM Rev.* **1990**, *32*, 474–476. [CrossRef]
62. Rohde, D.; Corcoran, J. MCMC methods for univariate exponential family models with intractable normalization constants. In Proceedings of the 2014 IEEE Workshop on Statistical Signal Processing (SSP), Gold Coast, Australia, 29 June–2 July 2014; pp. 356–359.

63. Barr, D.R.; Sherrill, E.T. Mean and variance of truncated normal distributions. *Am. Stat.* **1999**, *53*, 357–361.
64. Amendola, C.; Faugere, J.C.; Sturmfels, B. Moment Varieties of Gaussian Mixtures. *J. Algebr. Stat.* **2016**, *7*, 14–28. [CrossRef]
65. Fujisawa, H.; Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivar. Anal.* **2008**, *99*, 2053–2081. [CrossRef]
66. Nielsen, F.; Nock, R. Patch matching with polynomial exponential families and projective divergences. In Proceedings of the International Conference on Similarity Search and Applications, Tokyo, Japan, 24–26 October 2016, pp. 109–116.
67. Yang, Y.; Martin, R.; Bondell, H. Variational approximations using Fisher divergence. *arXiv* **2019**, arXiv:1905.05284.
68. Kostrikov, I.; Fergus, R.; Tompson, J.; Nachum, O. Offline reinforcement learning with Fisher divergence critic regularization. In Proceedings of the International Conference on Machine Learning (PMLR 2021), online, 7–8 June 2021; pp. 5774–5783.
69. Elkhalil, K.; Hasan, A.; Ding, J.; Farsiu, S.; Tarokh, V. Fisher Auto-Encoders. In Proceedings of the International Conference on Artificial Intelligence and Statistics (PMLR 2021), San Diego, CA, USA, 13–15 April 2021; pp. 352–360.
70. Améndola, C.; Engström, A.; Haase, C. Maximum number of modes of Gaussian mixtures. *Inf. Inference J. IMA* **2020**, *9*, 587–600. [CrossRef]
71. Aprausheva, N.; Mollaverdi, N.; Sorokin, S. Bounds for the number of modes of the simplest Gaussian mixture. *Pattern Recognit. Image Anal.* **2006**, *16*, 677–681. [CrossRef]
72. Aprausheva, N.; Sorokin, S. Exact equation of the boundary of unimodal and bimodal domains of a two-component Gaussian mixture. *Pattern Recognit. Image Anal.* **2013**, *23*, 341–347. [CrossRef]
73. Xiao, Y.; Shah, M.; Francis, S.; Arnold, D.L.; Arbel, T.; Collins, D.L. Optimal Gaussian mixture models of tissue intensities in brain MRI of patients with multiple-sclerosis. In Proceedings of the International Workshop on Machine Learning in Medical Imaging, Beijing, China, 20 September 2010; pp. 165–173.
74. Bilik, I.; Khomchuk, P. Minimum divergence approaches for robust classification of ground moving targets. *IEEE Trans. Aerosp. Electron. Syst.* **2012**, *48*, 581–603. [CrossRef]
75. Alippi, C.; Boracchi, G.; Carrera, D.; Roveri, M. Change Detection in Multivariate Datastreams: Likelihood and Detectability Loss. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016.
76. Eguchi, S.; Komori, O.; Kato, S. Projective power entropy and maximum Tsallis entropy distributions. *Entropy* **2011**, *13*, 1746–1764. [CrossRef]
77. Orjebin, E. A Recursive Formula for the Moments of a Truncated Univariate Normal Distribution. 2014, Unpublished note.
78. Del Castillo, J. The singly truncated normal distribution: a non-steep exponential family. *Ann. Inst. Stat. Math.* **1994**, *46*, 57–66. [CrossRef]