Composite distance based approach to von Mises mixture reduction

Mario Bukal[a,*], Ivan Marković[a], Ivan Petrović[a]

[a] *University of Zagreb, Faculty of Electrical Engineering and Computing,*
Unska 3, 10000 Zagreb, Croatia

Abstract

This paper presents a systematic approach for component number reduction in mixtures of exponential families, putting a special emphasis on the von Mises mixtures. We propose to formulate the problem as an optimization problem utilizing a new class of computationally tractable composite distance measures as cost functions, namely the composite Rényi α-divergences, which include the composite Kullback-Leibler distance as a special case. Furthermore, we prove that the composite divergence bounds from above the corresponding intractable Rényi α-divergence between a pair of mixtures. As a solution to the optimization problem we synthesize that two existing suboptimal solution strategies, the generalized k-means and a pairwise merging approach, are actually minimization methods for the composite distance measures. Moreover, in the present paper the existing joining algorithm is also extended for comparison purposes. The algorithms are implemented and their reduction results are compared and discussed on two examples of von Mises mixtures: a synthetic mixture and a real-world mixture used in people trajectory shape analysis.

Keywords: von Mises mixture, mixture component number reduction, Rényi α-divergence, composite distance measure, people trajectory shape analysis

[*] Corresponding author

Email addresses: mario.bukal@fer.hr (Mario Bukal), ivan.markovic@fer.hr (Ivan Marković), ivan.petrovic@fer.hr (Ivan Petrović)

1. Introduction

Many statistical and engineering problems [1–3] require modelling of complex multi-modal data, wherein mixture distributions became an inevitable tool. In this paper we draw attention to finite mixtures of a specific distribution on the unit circle, the von Mises distribution. Starting from 1918 and the seminal work of von Mises [4], where he investigated hypothesis on integrality of atomic weights of chemical elements, the proposed parametric density plays a pertinent role in directional statistics with wide range of applications in physics, biology, image analysis, neural science and medicine — confer monograms [5–7] and references therein.

Estimation of complex data by mixture distributions may lead to models with large or, in applications like target tracking, ever increasing number of components. In lack of efficient reduction procedures, such models become computationally intractable and lose their feasibility. Therefore, component number reduction in mixture models is an essential tool in many domains like image and multimedia indexing [8, 9], speech segmentation [10], and it is an indispensable part of any tracking system with mixtures of Gaussian [11–13] or von Mises distributions [3]. The subject matter is particularly relevant to the information fusion domain since it relates to the following challenging problems in multisensor data fusion [14]: data dimensionality, processing framework, and data association. These problems are related to component reduction by the fact that measurement data as quantity of interest can be preprocessed (compressed) prior to communicating it to other nodes (in a decentralized framework) or the fusion center, thus effectively saving on the communication bandwidth and power required for transmitting data. For example, consider the problem of people trajectory analysis with von Mises mixtures [2] in a distributed sensor networks where the mixtures might need to be communicated between the sensor nodes. Motivated by [2, 3], in this paper we study methods and respective algorithms for component number reduction in mixtures of von Mises distributions, but due to the general exposition of the subject in the framework of exponential family mixtures, the methods and findings easily extend to other examples like mixtures of Gaussian distributions, von Mises-Fisher distributions [5] etc.

Existing literature on mixture reduction schemes is mostly related to Gaussian mixture models. A reduction scheme for Gaussian mixtures in the context of Bayesian tracking systems in a cluttered environment, which succesively merges the closest pair of components and henceforth referred to as the *joining algorithm*, was proposed in [11]. The main drawback of the scheme is its local character, which gives no information about the global deviation of the reduced mixture from the original one. In [15] the mixture reduction was formulated as an optimization problem for the integral square difference cost function. A better suited distance measure between probability distributions is the Kullback-Leibler (KL) distance [16], but it lacks a closed form formula between mixtures, what makes it computationally inconvenient. Several concepts have been employed to circumvent this problem. A new distance measure between mixture distributions, based on the KL distance, which can be expressed analytically was derived in [17], and utilized to solve the mixture reduction problem. In [12] an upper bound for the KL distance was obtained and used as dissimilarity measure in a successive pairwise reduction of Gaussian mixtures — henceforth we refer to it as the *pairwise merging algorithm.* Unlike the joining algorithm, this procedure gives a control of the global deviation of the reduced mixture from the original one. Introducing the notion of Bregman information, the authors in [18] generalized the previously developed Gaussian mixture reduction concepts to arbitrary exponential family mixtures. Further development of these techniques for exponential family mixtures can be found in [19–24]. Finally, we mention the variational Bayesian approach [25, 26] as well as [27] as alternative concepts of mixture reduction developed for Gaussian mixtures.

Contributions of the present paper are as follows. Firstly, we formulate the problem of component number reduction in exponential family mixtures as an optimization problem utilizing a new class of composite distance measures as cost functions. These distance measures are constructed employing Rényi α-divergences as ground distances, and it is shown that the composite distance bounds the corresponding Rényi α-divergence from above (see Lemma 1 below). This inequality is very important since it provides an information on the global deviation of the reduced mixture from the original one measured by the Rényi α-divergence. Secondly, we synthesize previously developed reduction techniques [12, 18, 24] in the sense that they can all be interpreted as suboptimal solution strategies to the proposed optimization problem. For the purpose of computational complexity and accuracy comparisons, the joining algorithm is extended using the scaled symmetrized KL distance as a dissimilarity measure between mixture components. Thirdly, special attention is given to von Mises mixtures for which we present analytical expressions for solving the component number reduction problem and analyze them on two examples: a synthetic 100-component mixture with several dominant modes and a real-world mixture stemming from the work on people trajectory analysis in video data [2].

Outline of the paper is as follows. The general framework of exponential family mixtures is introduced in Section 2 together with a brief survey on distance measures between probability distributions and definition of composite distance measures. Section 3 presents the component number reduction in exponential family mixtures as a constrained optimization problem. In Section 4 we discuss two suboptimal solution strategies and additionally consider the joining algorithm. Numerical experiments on two examples of circular data are performed and obtained results are discussed in Section 5. Finally, Section 6 concludes the paper by outlining main achievements and commenting on possible extensions.

2. General background

In this section we introduce exponential family distributions and the von Mises distribution as their subclass, we recall the notion of finite mixture distributions and discuss variety of distance measures between probability distributions emphasizing on *composite distance measures* between mixtures.

2.1. Exponential family distributions

A parametric set of probability distributions defined on a sample space $\mathcal{X}$ and parametrized by the natural parameter $\theta \in \Theta \subset \mathbb{R}^d$ is called *exponential family* if their probability densities admit the following canonical representation

$$p_F(x;\theta) = \exp(T(x)\cdot\theta - F(\theta) + C(x)), \quad x \in \mathcal{X}. \tag{1}$$

Map $T : \mathcal{X} \to \mathbb{R}^d$ is called the minimal sufficient statistics, and functions F and C denote the log-normalizer (or log-partition) and the carrier measure, respectively. It can be proved that $\Theta = \mathrm{Dom}(F)$ is a nonempty convex set, and F is convex and unique up to an additive constant [28]. Moreover, if the exponential family is *regular* (i.e. Θ is open), then F is strictly convex and differentiable on Θ [18]. In further, the exponential family accompanied with the convex function F will be denoted by $\mathcal{E}_F$.

Many well known parametric distributions, like Gaussian, Poisson, Gamma, Dirichlet, etc., are exponential families [6]. For the reader's convenience, recall the simplest example of the univariate Gaussian distribution

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-(x-\mu)^2/2\sigma^2\right),$$

with standard parameters (μ,σ^2), which is an exponential family with natural parameter $\theta = (\mu/\sigma^2, 1/2\sigma^2) \in \mathbb{R}^2$, sufficient statistics $T(x) = (x, -x^2)$, log-normalizer $F(\theta) = \theta_1^2/4\theta_2 + \log(\pi/\theta_2)/2$, and $C(x) = 0$. Canonical parametrizations (1) for other exponential families can be found in [29], and in the sequel we focus on our study example — the von Mises distribution.

2.1.1. Von Mises distribution

The von Mises distribution is a probability distribution defined on the unit circle, or equivalently on the interval $[0, 2\pi)$, with density function given by

$$p(x; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\left\{\kappa \cos(x - \mu)\right\}, \quad 0 \leq x < 2\pi, \tag{2}$$

where $\mu \in [0, 2\pi)$ denotes the mean angle, $\kappa \geq 0$ is the concentration parameter, and I_0 is the modified Bessel function of the first kind and of order zero [5]. Recall, the modified Bessel function of the first kind and of order $n \in \mathbb{N}$ is defined by

$$I_n(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \exp(\kappa \cos\xi) \cos(n\xi) \, \mathrm{d}\xi. \tag{3}$$

In many ways von Mises distribution is considered as the circular analogue of the univariate Gaussian distribution: it is unimodal, symmetric around the mean angle μ, and the concentration parameter κ is analogous to the inverse of the variance. Furthermore, it is characterized by the maximum entropy principle in the sense that it maximizes the Boltzmann-Shannon entropy under prescribed circular mean [5].

From (2) it can be readily derived that von Mises distribution with standard parameters (μ, κ) is an exponential family parametrized by the natural parameter $\theta = (\kappa \cos\mu, \kappa \sin\mu) \in \Theta = \mathbb{R}^2$. The minimal sufficient statistics is the standard parametrization of the unit circle $T(x) = (\cos x, \sin x)$, the log-normalizer is given by

$$F(\theta) = \log\left(2\pi I_0\left(\sqrt{\theta_1^2 + \theta_2^2}\right)\right), \tag{4}$$

and the carrier measure is trivial, $C(x) = 0$.

2.2. Exponential family mixtures

A finite exponential family mixture distribution is a weighted normalized sum of distributions belonging to the same exponential family $\mathcal{E}_F$. Its density function is given by

$$p(x) = \sum_{i=1}^{K} w_i p_F(x; \theta_i), \quad x \in \mathcal{X}, \tag{5}$$

where K denotes the number of (different) parameters $\theta_i \in \Theta$ representing the mixture components and w_i are the corresponding weights which sum up to unity.

2.3. Distance measures

In order to analyze and approximate distributions one needs the notion of distance, either in proper or in generalized sense. There are numerous distance measures between distributions, but we will concentrate on those which are appropriate for finite mixtures, both from practical and theoretical aspects. Standard *integral distances* [15] between distributions do not take into account their key properties: the nonnegativity and normalization, what makes them often unsuitable in statistical analysis.

Statistically and information theoretically motivated distance measure is the *Kullback-Leibler (KL) distance* [16], also known as the Kullback-Leibler divergence or the relative entropy, defined by

$$D_{\mathrm{KL}}(p, q) = \int_{\mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)} \right) \mathrm{d}x.$$

The KL distance is a generalized distance functional, which is not symmetric nor it satisfies the triangle inequality, but it is positive definite, i.e. $D_{\mathrm{KL}}(p, q) \geq 0$ and $D_{\mathrm{KL}}(p, q) = 0$ only when $p = q$. It belongs to a wider class of distance measures called *f-divergences or Ali-Silvey distances* [30], which have numerous applications in statistics and information theory [31]. Another statistical and information theoretical class of generalized distances, particularly addressed in this paper, are *Rényi α-divergences* [32]

$$D_{R_\alpha}(p, q) = \frac{1}{\alpha - 1} \log \int_{\mathcal{X}} p(x)^{\alpha} q(x)^{1-\alpha} \mathrm{d}x$$

parametrized by the real parameter $\alpha \neq 1$. Using the l'Hospital's rule when $\alpha \to 1$, one obtains the KL distance at the limit. For further reading on generalized distance measures, their rich mathematical structure, wide range of information theoretical applications, as well as their intimate relations to exponential family distributions, we refer to a recent study in [33] and references therein.

Next, we present closed form expressions for the above defined distance measures between given exponential family distributions. Let $p = p_F(\cdot; \theta_p)$ and $q = p_F(\cdot; \theta_q) \in \mathcal{E}_F$, then straightforward calculations reveal the explicit formula

$$D_{\mathrm{KL}}(p, q) = B_F(\theta_q, \theta_p), \tag{6}$$

where B_F denotes the *Bregman divergence* [34] generated by the convex log-normalizing function F,

$$B_F(\theta_1, \theta_2) = F(\theta_1) - F(\theta_2) - \nabla F(\theta_2) \cdot (\theta_1 - \theta_2), \quad \theta_1, \theta_2 \in \Theta.$$

Equation (6) is valuable in the sense that translates the KL distance between exponential family densities to the Bregman divergence between the respective natural parameters, but in reversed order. Concerning Rényi α-divergences, they also admit a closed form expressions between exponential family distributions when $\alpha \in (0,1)$, and are given by

$$D_{R_\alpha}(p,q) = \frac{1}{1-\alpha} J_F^{(\alpha)}(\theta_p, \theta_q), \tag{7}$$

where $J_F^{(\alpha)}$ denotes the *Jensen α-divergence (or Burbea-Rao)* [35] generated by the convex function F,

$$J_F^{(\alpha)}(\theta_1, \theta_2) = \alpha F(\theta_1) + (1-\alpha) F(\theta_2) - F(\alpha\theta_1 + (1-\alpha)\theta_2), \quad \theta_1, \theta_2 \in \Theta.$$

Specifically, let p and q be von Mises distributions with standard parameters (μ_p, κ_p) and (μ_q, κ_q), respectively. Using the log-normalizing function (4) in (6), direct calculations reveal

$$D_{\mathrm{KL}}(p,q) = \log \frac{I_0(\kappa_q)}{I_0(\kappa_p)} + A(\kappa_p)(\kappa_p - \kappa_q \cos(\mu_p - \mu_q)), \tag{8}$$

where $A(\kappa) = I_1(\kappa)/I_0(\kappa)$. Similarly, for $\alpha \in (0,1)$ (7) gives Rényi α-divergences

$$D_{R_\alpha}(p,q) = \frac{\alpha}{1-\alpha} \log I_0(\kappa_p) + \log I_0(\kappa_q) + \frac{1}{\alpha - 1} \log I_0(\kappa_{pq}^{(\alpha)}), \tag{9}$$

where $\kappa_{pq}^{(\alpha)} = \sqrt{\alpha^2\kappa_p^2 + (1-\alpha)^2\kappa_q^2 + 2\alpha(1-\alpha)\kappa_p\kappa_q\cos(\mu_p - \mu_q)}$. Direct application of the above distances on mixture distributions does not give such closed form expressions depending on mixture parameters, which makes them typically impractical from computational viewpoint. However, the notion of composite (transport) distances, motivated by the optimal transportation theory [36], renders the aforementioned problem soluble.

2.3.1. Composite distance measures

Let $p = \sum_{i=1}^K w_i p_i$ and $q = \sum_{j=1}^L w'_j q_j$ be given exponential family mixtures, where p_i and q_j are abbreviations for $p_F(\cdot;\theta_i)$ and $p_F(\cdot;\theta'_j)$, respectively. Let $D(\cdot,\cdot)$ denote an f-divergence or Rényi α-divergence, then the *composite D-distance* [37] between mixtures p and q is defined by

$$d_D(p, q) = \inf_{u \in \Gamma(\boldsymbol{w}, \boldsymbol{w}')} \sum_{i,j} u_{ij} D(p_i, q_j), \tag{10}$$

where $\Gamma(\boldsymbol{w}, \boldsymbol{w}') = \{\boldsymbol{u} \in \mathbb{R}^{K \times L} \ : \ u_{ij} \geq 0, \ \sum_{i=1}^{K} u_{ij} = w'_j, \ \sum_{j=1}^{L} u_{ij} = w_i\}$ denotes the set of couplings between vectors $\boldsymbol{w}$ and $\boldsymbol{w}'$. The infimum in (10) is always achieved, since one optimizes a linear function over the compact set $\Gamma(\boldsymbol{w}, \boldsymbol{w}') \subset \mathbb{R}^{K \times L}$. In fact, computation of the composite distance d_D corresponds to solving a linear programming problem. It is easily seen that d_D is a generalized distance measure, i.e. $d_D(p, q) \geq 0$ and $d_D(p, q) = 0$ only if $p = q$. Main accomplishments of using the composite distances as distance measures between mixtures are twofold. First, their computation depends only on mixture parameters, and second, they give upper bounds on the corresponding ground distance between mixtures, as discussed in the following lemma.

Lemma 1. *Let p and q be finite mixtures, and let D be f-divergence or Rényi α-divergence with $\alpha \in (0, 1)$, then*

$$D(p, q) \leq d_D(p, q). \tag{11}$$

Proof. Let $p = \sum_{i=1}^{K} w_i p_i$ and $q = \sum_{j=1}^{L} w'_j q_j$ be given mixtures. The case when D is f-divergence has been proven in [37]. Thus, let $D = D_{R_\alpha}$ for some $\alpha \in (0, 1)$, and we write

$$D_{R_\alpha}(p, q) = \frac{1}{\alpha - 1} \log \int_{\mathcal{X}} p(x) \phi_\alpha \left(\frac{q(x)}{p(x)} \right) \mathrm{d}x$$

with $\phi_\alpha(u) = u^{1-\alpha}$. Since $(1 - \alpha) \in (0, 1)$, then ϕ_α is concave, and consequently $\psi_\alpha := -\phi_\alpha$ is convex. Following the lines of the proof for composite f-divergences in [37], for arbitrary $u \in \Gamma(\boldsymbol{w}, \boldsymbol{w}')$ we calculate

$$\begin{aligned} \int_{\mathcal{X}} p \, \phi_\alpha \left(\frac{q}{p} \right) \mathrm{d}x &= - \int_{\mathcal{X}} p \, \psi_\alpha \left(\frac{q}{p} \right) \mathrm{d}x \geq - \sum_{i,j} u_{ij} \int_{\mathcal{X}} p_i \, \psi_\alpha \left(\frac{q_j}{p_i} \right) \mathrm{d}x \\ &= \sum_{i,j} u_{ij} \int_{\mathcal{X}} p_i \, \phi_\alpha \left(\frac{q_j}{p_i} \right) \mathrm{d}x. \end{aligned} \tag{12}$$

Taking the logarithm of (12), using the Jensen's inequality, and multiplying by $1/(\alpha - 1) < 0$ we obtain

$$\frac{1}{\alpha - 1} \log \int_{\mathcal{X}} p \, \phi_\alpha \left(\frac{q}{p} \right) \mathrm{d}x \leq \sum_{i,j} u_{ij} \frac{1}{\alpha - 1} \log \int_{\mathcal{X}} p_i \, \phi_\alpha \left(\frac{q_j}{p_i} \right) \mathrm{d}x.$$

The left-hand side in the last inequality is exactly $D_{R_\alpha}(p, q)$. Since the inequality holds for any $u \in \Gamma(\boldsymbol{w}, \boldsymbol{w'})$, we can take the infimum over $\Gamma(\boldsymbol{w}, \boldsymbol{w'})$ on the right-hand side, which finishes the proof. □

If D is the KL distance, the composite KL distance $d_{\mathrm{KL}}(p, q)$ can be interpreted as the total cost of coding data generated by p under the model q [17]. Note as well that in the above lemma there is no requirement on p and q being exponential family mixtures and that the statement holds true for general finite mixtures.

3. Problem formulation

Having defined suitable distance measures from the previous section, we formulate the problem of reduction of the number of components, described in Section 1, as follows. Let $p = \sum_{i=1}^{K} w_i p_i$ be the given starting exponential family mixture and let D denote the chosen ground distance, the KL or Renyi α-divergence with $\alpha \in (0, 1)$. The optimization problem

$$\min_{q' \in \mathcal{M}_L} d_D(p, q') \tag{13}$$

aims to find a mixture q' having at most L components, which is the best approximation of p with respect to the composite D-distance. If we denote

$$D_F^{(\alpha)}(\theta_i, \theta'_j) = \begin{cases} \dfrac{1}{1-\alpha} J_F^{(\alpha)}(\theta_i, \theta'_j), & \alpha \in (0, 1), \\ B_F(\theta'_j, \theta_i), & \alpha = 1, \end{cases} \tag{14}$$

then according to the definition of composite distance and having in mind explicit formulae (6)–(7), the above problem amounts to a constrained (nonlinear) optimization problem

$$\min_{\boldsymbol{\theta'},\, \boldsymbol{w'},\, \boldsymbol{u}} \sum_{i=1}^{K} \sum_{j=1}^{L} u_{ij} D_F^{(\alpha)}(\theta_i, \theta'_j) \tag{15}$$

over unknown natural parameters $\boldsymbol{\theta'} = (\theta'_1, \ldots, \theta'_L) \in \Theta^L$, its weights $\boldsymbol{w'} = (w'_1, \ldots, w'_L) \in [0, 1]^L$ and optimal couplings $\boldsymbol{u} \in \Gamma(\boldsymbol{w}, \boldsymbol{w'})$. Associated constraints are given by linear equations

$$\sum_{j=1}^{L} u_{ij} = w_i,\ i = 1, \ldots, K; \quad \sum_{i=1}^{K} u_{ij} = w'_j,\ j = 1, \ldots, L; \quad \sum_{j=1}^{L} w'_j = 1, \tag{16}$$

and the nonnegativity conditions $u_{ij} \geq 0$ for all $i = 1, \dots, K$ and $j = 1, \dots, L$. The last two equations in (16) imply $\sum_{i,j} u_{ij} = 1$, which is trivially fulfilled by the first equation in (16) due to the normalizing condition $\sum_{i=1}^{K} w_i = 1$. Hence, the only equality constraints left are

$$\sum_{j=1}^{L} u_{ij} = w_i, \quad i = 1, \dots, K. \tag{17}$$

Due to nonlinearities in natural parameters, globally optimal solution to the problem (15) and (17) seems to be out of reach. Instead, we will present two suboptimal solution strategies in the forthcoming section, but first we derive necessary conditions for optimality.

3.1. Necessary conditions for local minimizers

Fixing $\alpha \in (0, 1]$, the Lagrange function of the optimization problem (15) with constraints (17) equals

$$L(\boldsymbol{\theta}', \boldsymbol{w}', \boldsymbol{z}, \boldsymbol{\lambda}) = \sum_{i,j} z_{ij}^2 D_F^{(\alpha)}(\theta_i, \theta'_j) + \sum_{i=1}^{K} \lambda_i \Big(w_i - \sum_{j=1}^{L} z_{ij}^2 \Big) \tag{18}$$

where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K)$ are Lagrange multipliers and the change of variables $u_{ij} = z_{ij}^2$ substitutes the nonnegativity conditions $u_{ij} \geq 0$.

Taking gradients with respect to θ'_j and equating with zero yields the set of equations

$$w'_j \nabla F(\theta'_j) = \sum_{i=1}^{K} z_{ij}^2 \nabla F(\alpha\theta_i + (1-\alpha)\theta'_j), \quad j = 1, \dots, L. \tag{19}$$

Partial derivatives with respect to z_{ij} give conditions

$$z_{ij}(D_F^{(\alpha)}(\theta_i, \theta'_j) - \lambda_i) = 0, \quad i = 1, \dots, K; \ j = 1, \dots, L. \tag{20}$$

Now we determine $\boldsymbol{\lambda}$. Let $\boldsymbol{z} = (z_{ij})$ be a stationary point of (18). If $z_{ij} \neq 0$ for some i and j, then according to (20)

$$\lambda_i = D_F^{(\alpha)}(\theta_i, \theta'_j).$$

The same holds true for fixed i and arbitrary j such that $z_{ij} \neq 0$, hence λ_i is well defined only when all θ'_j (with j such that $z_{ij} \neq 0$) lie on the same $D_F^{(\alpha)}$-distance from θ_i. Therefore,

$$\lambda_i = \min_{j \in N_i} D_F^{(\alpha)}(\theta_i, \theta'_j), \quad i = 1, \dots, K,$$

where $N_i = \{j \in \{1, \ldots, L\} \ : \ z_{ij} \neq 0\}$. So far, for given stationary $\boldsymbol{z}$ we have equations for θ'_j (given by (19)), and prescribed relations for the corresponding weights $w'_j = \sum_{i=1}^{K} z_{ij}^2$. One still needs to determine optimal $\boldsymbol{z} = (z_{ij})$. Applying conditions (17) and the above discussion about $\boldsymbol{\lambda}$ reveals that $\boldsymbol{z} = (z_{ij})$, defined by

$$z_{ij} = \begin{cases} \sqrt{w_i}, & j = \operatorname*{arg\,min}_{l \in \{1,\ldots,L\}} D_F^{(\alpha)}(\theta_i, \theta'_l), \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

is a stationary point. However, this definition depends on the stationary $\boldsymbol{\theta}'$. Consequently, we will need to employ iterative procedures to solve the stationary problems (19) and (21).

In the case of univariate Gaussian mixtures, necessary conditions (19), in terms of standard parameters, reduce to the following system of algebraic equations:

$$\begin{aligned} w'_j \mu'_j &= \sum_{i=1}^{K} z_{ij}^2 \frac{\alpha \sigma_{ij}^2 \mu_i + (1-\alpha)\mu'_j}{\alpha \sigma_{ij}^2 + 1 - \alpha}, \\ w'_j (\sigma'^2_j + \mu'^2_j) &= \sum_{i=1}^{K} z_{ij}^2 \Big(\frac{\sigma'^2_j}{\alpha \sigma_{ij}^2 + 1 - \alpha} + \Big(\frac{\alpha \sigma_{ij}^2 \mu_i + (1-\alpha)\mu'_j}{\alpha \sigma_{ij}^2 + 1 - \alpha} \Big)^2 \Big), \end{aligned}$$

where $\sigma_{ij} = \sigma'_j / \sigma_i$. Note that if $\alpha = 1$, they become explicit formulae for mean values μ'_j and variances σ'^2_j. Similarly holds true for multivariate Gaussian mixtures (cf. [11, Eq. (3)]).

In general, system (19) is heavily nonlinear, for instance in the von Mises case it involves the ratio of Bessel functions (see below). However, if $\alpha = 1$, the right hand side in (19) depends only on coupling $\boldsymbol{z}$ and the original set of parameters. A unique solution is asserted by strict convexity of the function F and can be obtained by applying the Newton method. If $\alpha \in (0, 1)$, solving system (19) requires an iterative convex-concave optimization scheme, which eventually converges to the unique solution of (19) — see [24] for a detailed analysis of such systems.

3.2. The case of von Mises mixtures

In light of the above discussion on exponential families, in this subsection we highlight the problem of component number reduction (15) in case of the von Mises mixtures and eventually express the necessary conditions (19) and (21) in terms of the standard parameters of von Mises mixtures. Using the relation $\nabla F(\theta) = A(\kappa)(\cos\mu, \sin\mu)$ between the natural and standard parameters, equation (19) translates (componentwise) to

$$w_j' A(\kappa_j') \cos\mu_j' = \sum_{i=1}^{K} z_{ij}^2 A(\kappa_{ij}^{(\alpha)}) \cos(\mu_{ij}^{(\alpha)}), \quad j = 1, \ldots, L, \tag{22}$$

$$w_j' A(\kappa_j') \sin\mu_j' = \sum_{i=1}^{K} z_{ij}^2 A(\kappa_{ij}^{(\alpha)}) \sin(\mu_{ij}^{(\alpha)}), \quad j = 1, \ldots, L, \tag{23}$$

where $\kappa_{ij}^{(\alpha)}$ and $\mu_{ij}^{(\alpha)}$ are defined by

$$\begin{aligned}
\kappa_{ij}^{(\alpha)} &= \sqrt{\alpha^2\kappa_i^2 + (1-\alpha)^2\kappa_j'^2 + 2\alpha(1-\alpha)\kappa_i\kappa_j' \cos(\mu_i - \mu_j')}, \\
\tan\mu_{ij}^{(\alpha)} &= \frac{\alpha\kappa_i \sin\mu_i + (1-\alpha)\kappa_j' \sin\mu_j'}{\alpha\kappa_i \cos\mu_i + (1-\alpha)\kappa_j' \cos\mu_j'}.
\end{aligned}$$

Dividing (23) by (22) we implicitly express μ_j', while squaring (22), (23) and summing them yields expressions for κ_j' as follows:

$$\begin{aligned}
\tan\mu_j' &= \frac{\sum_{i=1}^{K} z_{ij}^2 A(\kappa_{ij}^{(\alpha)}) \sin\mu_{ij}^{(\alpha)}}{\sum_{i=1}^{K} z_{ij}^2 A(\kappa_{ij}^{(\alpha)}) \cos\mu_{ij}^{(\alpha)}}, \quad j = 1, \ldots, L, \\
{w_j}'^2 A^2(\kappa_j') &= \sum_{i=1}^{K} z_{ij}^4 A^2(\kappa_{ij}^{(\alpha)}) \qquad\qquad\qquad\qquad (24) \\
&\quad + 2 \sum_{\substack{i,k=1 \\ i<k}}^{K} z_{ij}^2 z_{kj}^2 A(\kappa_{ij}^{(\alpha)}) A(\kappa_{kj}^{(\alpha)}) \cos(\mu_{ij}^{(\alpha)} - \mu_{kj}^{(\alpha)}), \quad j = 1, \ldots, L.
\end{aligned}$$

Note again that for $\alpha = 1$ the right hand side in (24) depends only on the coupling $\boldsymbol{z}$ and the known parameters μ_i and κ_i, while $\alpha \in (0, 1)$ makes (24) a heavily nonlinear problem in μ_j' and κ_j', which can be solved by the previously mentioned iterative convex-concave optimization [24]. Using (8) or (9) we also obtain explicit formulae for the stationary couplings $\boldsymbol{z}$ in (21), which, in other words, tell us which components we need to merge, while (24) tells us exactly how to calculate the merging.

4. Component reduction schemes

In this section we present three different approaches for solving the component number reduction problem. The first two approaches: *(i) generalized* *k-means clustering*, and *(ii) a gradual pairwise merging scheme*, present solution strategies which aim to solve the optimization problem (15), i.e. to minimize the composite distance between the original and the reduced mixture. The third approach, *the joining algorithm*, is a heuristic reduction scheme which successively merges a pair of components with the lowest mutual distance, and is considered here for comparison purposes.

4.1. Minimizing the composite distance

(i) Generalized k-means clustering

A natural iterative procedure for solving (19) and (21) is the *Lloyd's algorithm* [38] or the *generalized k-means clustering.* Depending on the choice of parameter $\alpha \in (0,1]$, the clustering algorithms are known in the literature as *Bregman clustering* for $\alpha = 1$ [18], and *α-Jensen (or Burbea-Rao) clustering* when $\alpha \in (0,1)$ [24]. However, the fact that these algorithms optimize the composite distance between mixtures and consequently provide the upper bound on the ground distance (cf. Lemma 1), is to the best of our knowledge novel.

Let $p = \sum_{i=1}^{K} w_i p_i$ denote the original mixture with natural parameters $\theta_i,\ i = 1, \ldots, K$, let the number of components of the reduced mixture L be given, and let $q = \sum_{j=1}^{L} w'_j q_j$ denote the reduced mixture represented by natural parameters $\theta'_j,\ j = 1, \ldots, L$, which needs to be determined. The first step of the algorithm requires initialization of parameters θ'_j. This can be done for example by taking random L points from the set $\{\theta_i\}$, or in a genuine way discussed below. For the moment, assume we are given an initial set of parameters $\{\theta'^{(0)}_j\}$. The generalized k-means clustering is a two step iterative scheme which consists of the *assignment step* and and the *recalculation of parameters.*

First, in the assignment step the coupling $\boldsymbol{z}^{(1)} = (z^{(1)}_{ij})$ is computed according to (21) using initial parameters $\theta'^{(0)}_j$:

$$z^{(1)}_{ij} = \begin{cases} \sqrt{w_i}, & j = \underset{l \in \{1,\ldots,L\}}{\arg\min}\, D^{(\alpha)}_F(\theta_i, \theta'^{(0)}_l), \\ 0, & \text{otherwise.} \end{cases}$$

Definition of $\boldsymbol{z}^{(1)}$ gives a set of clusters $\mathcal{C}^{(1)} = \{C^{(1)}_j : j = 1, \ldots, L\}$ which partition the original set $\{\theta_i\}$, such that $\theta_i \in C^{(1)}_j$ exactly when $z^{(1)}_{ij} \neq 0$. In the second step — *recalculation of parameters* — for given $\boldsymbol{z}^{(1)}$, equations in

(19) need to be solved to obtain new parameters $\theta_j'^{(1)}$:

$$w_j'^{(1)} \nabla F\big(\theta_j'^{(1)}\big) = \sum_{i=1}^{K} (z_{ij}^{(1)})^2 \nabla F\big(\alpha\theta_i + (1-\alpha)\theta_j'^{(1)}\big), \quad j = 1, \ldots, L, \qquad (25)$$

where $w_j'^{(1)} = \sum_{i=1}^{K} (z_{ij}^{(1)})^2$. In the next iteration, coupling $\boldsymbol{z}^{(2)}$ is calculated using parameters $\theta_j'^{(1)}$ from the first step, and the algorithm continues with calculation of new parameters $\theta_j'^{(2)}$. The above described procedure iterates between those two steps until $\boldsymbol{z}^{(k+1)} = \boldsymbol{z}^{(k)}$ for some $k \geq 1$, and the stopping criteria is guaranteed by the finitness of the set of all possible couplings. The scheme monotonically decreases the cost functional, i.e. the composite distance, in (15) (cf. [18]) for $k \in \mathbb{N}$

$$\begin{aligned}\sum_{i,j} (z_{ij}^{(k)})^2 D_F^{(\alpha)}(\theta_i, \theta_j'^{(k)}) &\geq \sum_{i,j} (z_{ij}^{(k+1)})^2 D_F^{(\alpha)}(\theta_i, \theta_j'^{(k)}) \\ &\geq \sum_{i,j} (z_{ij}^{(k+1)})^2 D_F^{(\alpha)}(\theta_i, \theta_j'^{(k+1)}),\end{aligned}$$

where these inequalities follow from the assignment step and recalculation of parameters, respectively. The obtained parameters $\theta_j' = \theta_j'^{(k)}$ and the coupling $\boldsymbol{z} = \boldsymbol{z}^{(k)}$, for some $k \in \mathbb{N}$, satisfy the necessary conditions (19) and (21), hence, they are a local minimum of (15), and the mixture $q = \sum_{j=1}^{L} w_j' q_j$ is a suboptimal solution to the component number reduction problem. Nothing can be said about the global optimality of the obtained solutions. Algorithm 1 summarizes the above presented scheme.

Due to their local optimality, evident drawback of the k-means scheme is strong dependence on initial conditions. A random choice of the initial guess might yield suboptimal solution very far from the globally optimal. In [39] a clever way to construct an initial guess was proposed which typically gives better results. The construction consists of repetitive sampling of a parameter from the set of original parameters $\{\theta_i\}$ according to a non-uniform discrete probability distribution defined by

$$P(\theta_i) = \frac{w_i \min_{\theta_j^{(0)} \in \mathcal{Q}_0} D_F^{(\alpha)}(\theta_i, \theta_j^{(0)})}{\sum_{i'=1}^{K} w_{i'} \min_{\theta_j^{(0)} \in \mathcal{Q}_0} D_F^{(\alpha)}(\theta_{i'}, \theta_j^{(0)})}, \quad i = 1, \ldots, K, \qquad (26)$$

where $\mathcal{Q}_0$ denotes the set of already sampled parameters. From (26) we see that sampling probabilities are proportional to the parameter weights and

Algorithm 1 Generalized k-means clustering

```
Require: Component parameters $\{\theta_i : i = 1, \ldots, K\} \subset \Theta$ with corresponding weights $\{w_i : i = 1, \ldots, K\}$, distance parameter $\alpha \in (0, 1]$
Ensure: Reduced component parameters $\{\theta'_j : j = 1, \ldots, L\} \subset \Theta$ with corresponding weights $\{w'_j : j = 1, \ldots, L\}$.

Parameter initialization: $\{\theta'^{(0)}_j : j = 1, \ldots, L\} \subset \Theta$
$k = 1$
$\boldsymbol{z}^{(0)} = \emptyset$
repeat
    # Assignment step
    for $i = 1, \ldots, K$ do
        $j^* = \operatorname*{arg\,min}_{j=1,\ldots,L} D_F^{(\alpha)}(\theta_i, \theta'^{(k-1)}_j)$
        $z^{(k)}_{ij} = (j == j^*)\ ?\ \sqrt{w_i}\ :\ 0,\ \ j = 1, \ldots, L$
    end for
    # Recalculation step
    for $j = 1, \ldots, L$ do
        $w'^{(k)}_j = \sum_{i=1}^{K} (z^{(k)}_{ij})^2$
        Solve (19) for $\theta'^{(k)}_j$
    end for
    $k = k + 1$
until $\boldsymbol{z}^{(k-1)} = \boldsymbol{z}^{(k-2)}$
```

to the distance from $\mathcal{Q}_0$. In that way the construction takes into account importance of each parameter θ_i and its relative position to the set of already sampled parameters. Parameters that are already sampled have probability zero for being sampled again. In Algorithm 2 we summarize the initialization procedure.

(ii) Pairwise merging scheme

Unlike the previous strategy which minimizes the composite distance by iterative clustering of the original set of parameters, in this section we discuss another scheme which gradually reduces the number of components by pairwise merging such that in each step the pair of components of the lowest cost (see (27) below) is merged. The scheme was first introduced in [12] for the case of Gaussian mixtures using the "Kullback-Leibler cost" ($\alpha = 1$ in (27)). Here we extend the idea to the case of exponential family mixtures and "Rényi cost functions". Moreover, we discuss how this scheme

Algorithm 2 Parameter initialization

```
Require: Component parameters $\mathcal{P} = \{\theta_i : i = 1, \dots, K\} \subset \Theta$ with corresponding weights $\{w_i : i = 1, \dots, K\}$, distance parameter $\alpha \in (0, 1]$
Ensure: Initial guess $\mathcal{Q}_0 = \{\theta_j'^{(0)} : j = 1, \dots, L\} \subset \Theta$ for $k$-means iterations

$\mathcal{Q}_0 \leftarrow \emptyset$
Sample a point $\theta_i \in \mathcal{P}$ according to the discrete (non-uniform) distribution $\boldsymbol{w} = (w_1, \dots, w_K)$
$\mathcal{Q}_0 \leftarrow \{\theta_i\}$
while $|\mathcal{Q}_0| < L$ do
    Sample $\theta_i \in \mathcal{P}$ according to (26), $\mathcal{Q}_0 \leftarrow \mathcal{Q}_0 \cup \{\theta_i\}$
end while
```

is related to minimization of the composite distances.

Since in this scheme each subsequent step follows the same lines, we will concentrate on the first step of the procedure. Again let $p = \sum_{i=1}^{K} w_i p_i$ denote the original mixture, which we now want to reduce to a $(K-1)$-component mixture q. Naive application of the k-means algorithm would first require to sample $K-1$ elements from the original set $\{\theta_i\}$. Then the assignment step would clearly result a simple clustering $\mathcal{C}^{(1)}$ where $K-1$ original elements are assigned to itself, and only one (the non-sampled) element is assigned elsewhere. Instead of proceeding with k-means, the idea is to consider all possible simple clustrings $\mathcal{C}^{(1)}$, which partition the original set of parameters into $K-1$ subsets, and are described by couplings $\boldsymbol{z} = (z_{ij}) \in \mathbb{R}^{K \times (K-1)}$ of the form (21). There are exactly $K(K-1)/2$ such couplings, and for each we can compute the corresponding cost in (15). Since each coupling fixes $K-2$ parameters (components) from p and exactly two parameters are to be merged, the optimization problem (15) simplifies to

$$\min_{\theta' \in \Theta} \min_{i,k} \left\{ w_i D_F^{(\alpha)}(\theta_i, \theta') + w_k D_F^{(\alpha)}(\theta_k, \theta') \right\}. \tag{27}$$

For each pair of indices $(i, k) \in \{1, \dots, K\}^2$, merged parameter θ' must satisfy the necessary conditions

$$w' \nabla F(\theta') = w_i \nabla F(\alpha\theta_i + (1-\alpha)\theta') + w_k \nabla F(\alpha\theta_k + (1-\alpha)\theta'), \tag{28}$$

with $w' = w_i + w_k$. A pair $(i^*, k^*) \in \{1, \dots, K\}^2$ with the minimal cost in (27) will be merged to reduce one component of the mixture. The procedure is repeated until the desired number of components L is reached. The pairwise merging scheme gains global optimality in each reduction step. However, that does not guarantee the overall global optimality of the reduction procedure after $N - L$ steps. Moreover, due to additional nonlinearity in the right hand side of (28) in case of $\alpha \in (0, 1)$, the obtained set of parameters $\{\theta'_j\}$ does not satisfy the necessary conditions (19). Hence, in order to improve the reduced parameters, the final coupling $\boldsymbol{z} \in \mathbb{R}^{K \times L}$ between the original and reduced mixture has to be reconstructed from the history of all consecutive pairwise couplings, and the convex-concave optimization needs to be performed to solve (19) taking the above obtained $\{\theta'_j\}$ as an initial guess.

4.2. The joining algorithm

Here we outline a heuristic approach, the joining algorithm, for solving the problem of component number reduction in finite mixtures. First proposed in [40], and later revised and extended in [11], this was one of the first approaches for component number reduction in mixtures of Gaussian distributions. The joining algorithm relies on finding a pair of components with the smallest mutual distance, and then merging that pair into one component.

In Section 2.3 we mentioned several distance measures between probability distributions, which, in order to be utilized here, need to be customized for weighted (unnormalized) distributions, and additionaly symmetrized. For the purpose, we consider *the scaled symmetrized KL distance* [41], defined by

$$D_{s\mathrm{KL}}(w_i p_i, w'_j q_j) = \frac{1}{2}(w_i D_{\mathrm{KL}}(p_i, q_j) + w'_j D_{\mathrm{KL}}(q_j, p_i)) + \frac{1}{2}(w_i - w'_j) \log \frac{w_i}{w'_j}, \tag{29}$$

where $w_i,\ w'_j > 0$. Scaled symmetrized Rényi α-divergences [32] could be also introduced at this point, but they are not appropriate since they neglect the respective weights of components, namely it can be shown that $D_{sR_\alpha}(w_i p_i, w'_j q_j) = (D_{R_\alpha}(p_i, q_j) + D_{R_\alpha}(q_j, p_i))/2$. The final goal of the joining algorithm is then to find the pair of components with the smallest mutual $D_{s\mathrm{KL}}$ distance and calculate the corresponding merged parameters according to (28) with $\alpha = 1$. The process is outlined in Algorithm 3.

Using an implementation which stores mutual distances between those components that do not change between two reduction steps, it is stright-forward to calculate that the number of distance evaluations is of order $\mathcal{O}(K^2 - L^2/2)$, while the number of comparisons in finding the minimum

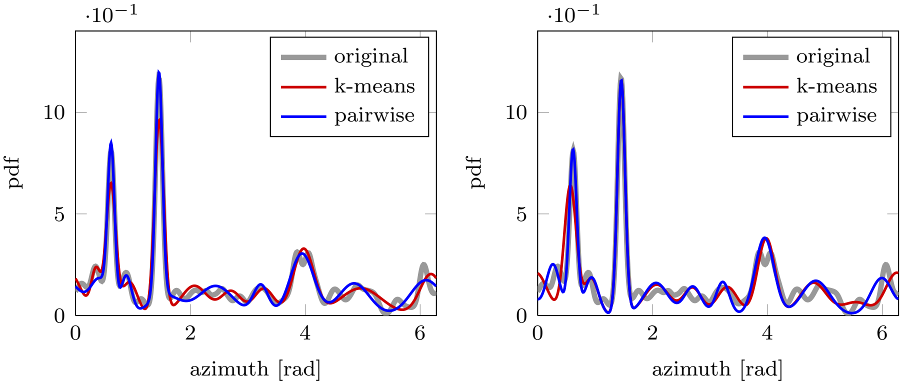
Algorithm 3 The joining algorithm

```
Require: Component parameters $\mathcal{P} = \{\theta_i : i = 1, \dots, K\} \subset \Theta$ with corresponding weights $\{w_i : i = 1, \dots, K\}$.
Ensure: Reduced component parameters $\mathcal{Q} = \{\theta'_j : j = 1, \dots, L\} \subset \Theta$ with corresponding weights $\{w'_j : j = 1, \dots, L\}$.
$d_{ij} \leftarrow \emptyset$
while $|\mathcal{P}| > L$ do
    for all $\{i, j\}$ with unknown $d_{ij}$ do
        $d_{ij} \leftarrow D_{sKL}(w_i p_i, w_j p_j)$
    end for
    $\{i^*, j^*\} = \arg\min_{i,j} d_{ij}$
    To get $\{\theta', w'\}$ solve (19) for $\{\theta_{i^*}, w_{i^*}\}$, $\{\theta_{j^*}, w_{j^*}\}$ and $\alpha = 1$
    $\mathcal{P} \leftarrow \mathcal{P} \setminus \{\theta_{i^*}, \theta_{j^*}\}$
    $\mathcal{P} \leftarrow \mathcal{P} \cup \{\theta'\}$
end while
$\mathcal{Q} \leftarrow \mathcal{P}$
```

distance value is $\mathcal{O}(K^3 - L^3)$ [11]. Hence, the algoritham is computationally demanding for K big and L relatively small. A simplification of the joining algorithm can be done by first sorting the components according to their weights and then calculating the distance between the component with the smallest weight and all other components of the mixture. Once the components with the smallest distance are merged, the new component is inserted according to its resulting weight. The process is repeated until the required number of components is reached. The idea behind is that in each step we merge the component which brings the least information to the mixture. This approach known as the West's algorithm is one of the computationally most efficient and it was proposed in [42] for component number reduction of mixtures of Gaussian distributions.

5. Results and discussion for von Mises mixtures

To test and compare the reduction algorithms for von Mises mixtures we utilized two examples. The first is a synthetic mixture consisting of 100 components chosen in a random manner, but with two components having a dominant weight in order to ensure a couple of dominant modes in the mixture. The second mixture is a real-world example steming from a people trajectory analysis dataset [2].



(a) $\alpha = 1$

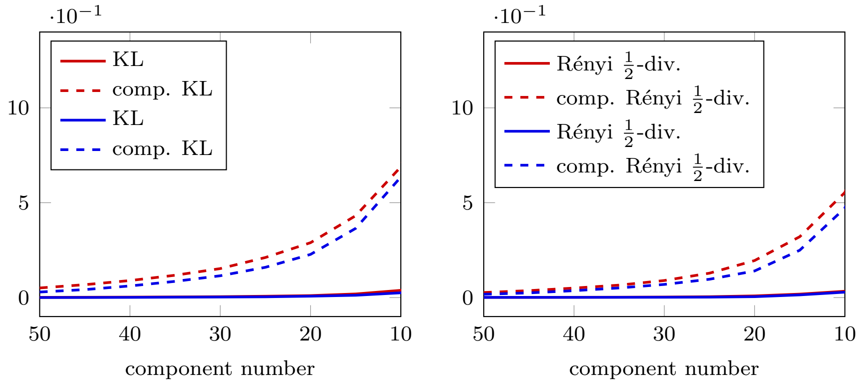
(b) $\alpha = 0.5$

Figure 1: k-means and pairwise merging of the synthetic mixture to 10 components by minimization of composite Rényi α-divergence.

5.1. Synthetic mixture

In this example a synthetic 100-component von Mises mixture constructed in a random manner, but with two dominant modes, needs to be reduced to a 10-component mixture.

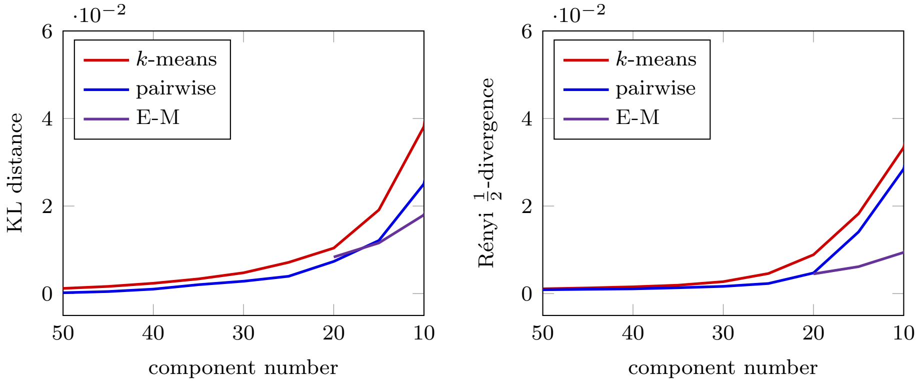
In order to perform the component number reduction, as proposed in Section 3, we first choose an appropriate distance measure. According to the literature, standard choices for the ground distance measure are the KL and the Rényi $\frac{1}{2}$-divergence (Bhattacharyya distance), which then lead to the composite KL distance and the composite Rényi $\frac{1}{2}$-divergence ($\alpha = 1$ and $\alpha = \frac{1}{2}$ in (14)), utilized in this particular example. Fig. 1 shows the original mixture and its reductions to 10 components by applying the pairwise and k-means algorithms for minimizing: (a) the composite KL distance and (b) the composite Rényi $\frac{1}{2}$-divergence. In Figs. 2 and 3 we analyze gradual reduction of the number of components, in steps of 5 components starting with 50 and ending with 10 components, using the k-means algorithm and the pairwise merging scheme for both composite distance measures. In Fig. 2 we show: (a) the optimized composite KL distance and (b) the composite Rényi $\frac{1}{2}$-divergence, along with the corresponding ground distances, respectively. Fig. 2 shows that, indeed, the composite distance is an upper bound on the ground distance and, moreover, that the pairwise merging, which is the more exhaustive scheme, shows better performance yielding a suboptimal solution closer to the global minimum. Figure 3 resolves the results from Fig. 2 in more detail by depicting only the corresponding ground



(a) $\alpha = 1$

(b) $\alpha = 0.5$

Figure 2: Composite distance and final Rényi α-divergence for the minimization via k-means (red) and pairwise (blue) methods.



(a) $\alpha = 1$

(b) $\alpha = 0.5$

Figure 3: Reduction results with respect to the corresponding Rényi α-divergence for the composite distance minimization methods along with the E-M with 5000 samples.

distances between the original and reduced mixtures. Again we can see that on average the pairwise merging approach to composite distance minimization showed better results. Also, for comparison purposes and a reduction quality assessment, the results of the E-M algorithm using 5000 samples are presented.

For all the algorithms and for each component number the reduction was

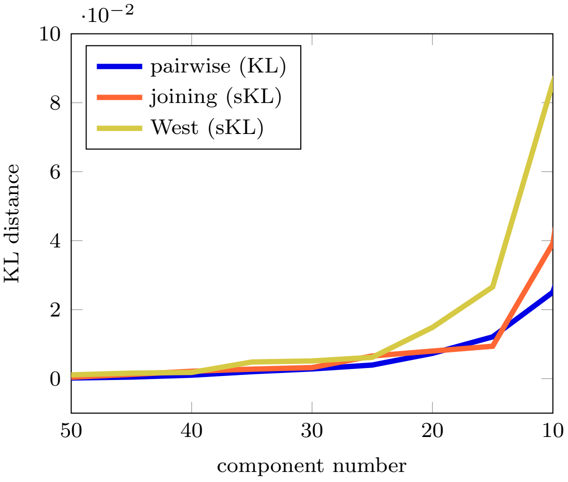


Figure 4: KL distance of the component reduction with the joining algorithm, West's algorithm, and the pairwise composite distance minimization.

repeated 50 times in order to get an average performance of the reduction and the execution time. We find it important for reliable measurement of not just the performance of algorithms with stochastic elements, like the k-means and E-M due to initialization and random sampling, but also for the execution time of all the algorithms.

Furthermore, for this example we also employed the heuristic approaches of joining and West for the component number reduction presented in Section 4.2, where we used the symmetrized KL distance for similarity comparison between components. Obtained results for gradual component number reduction as above, along with the pairwise minimization of composite KL distance, are compared in Fig. 4, from which we can see that for most components and notably 10 components the pairwise merging minimization keeps the best performance measured by the KL distance. Although the West's algorithm showed the lowest performance, it should not be dismissed lightly since it has the lowest execution time as discussed in the sequel.

Table 1 shows the execution time of the algorithms utilized for the reduction of the synthetic mixture. From the table we can draw several conclusions. As expected, the computationally least expensive algorithm is the West's algorithm, then followed by the k-means algorithm. The most expensive reduction scheme is the pairwise merging approach since in each step it requires solving a nonlinear problem with the nonlinearity being the ratio of Bessel functions. Furthermore, the methods utilizing the Rényi $\frac{1}{2}$-divergence

Table 1: Reduction time of tested algorithms for the synthetic mixture of von Mises distributions.

		Reduction time[s]					
		comp. KL		comp. Rényi $\frac{1}{2}$-div.		symm. KL	
		k-means	pairwise	k-means	pairwise	joining	West
Comp.	30	11.33	50.88	12.31	177.81	10.30	3.65
	20	7.35	51.29	8.21	180.65	10.36	3.78
	10	4.09	51.49	4.50	182.44	10.44	3.86

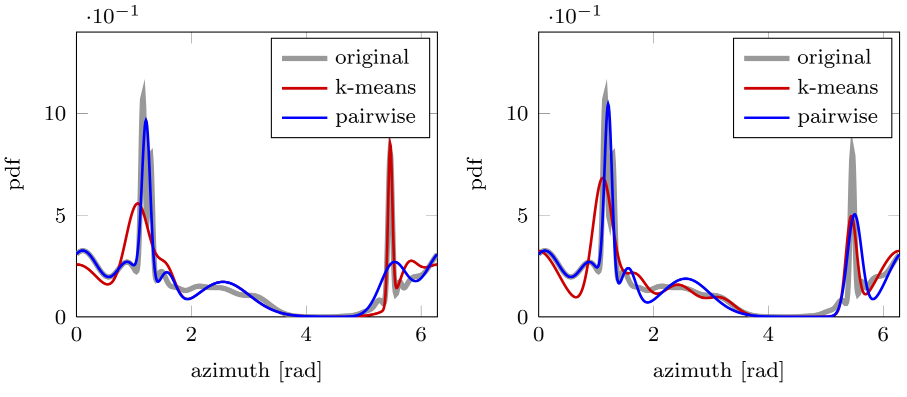
Results were obtained on an Intel® Core™ i7 CPU running at 1.6 GHz

are even more expensive than their KL counterparts since they additionally have to employ the iterative convex-concave optimization scheme to calculate the components parameters. For comparison purposes, we ran the pairwise merging method minimizing the composite KL distance on a 100-component univariate Gaussian mixture. For reductions to 30, 20, and 10 components the reduction time was 3.97 s, 3.99 s, and 4.05 s, respectively. This example enlightens the computational demand of working with von Mises mixtures due to special functions and numerical procedures involved.

5.2. People trajectory dataset mixture

Second example is a real-world mixture coming from people trajectory shape analysis. Raw data is taken from the webpage of authors from [2] and original 18-component von Mises mixture was obtained by employing the standard E-M algorithm to provided samples. The task is to reduce the original mixture to a 6-component mixture of von Mises distributions.

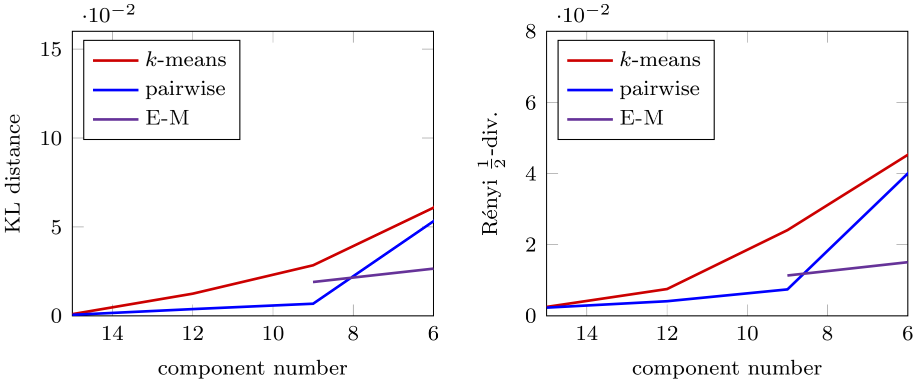
Here, the strategy was similar as for the synthetic mixture. First, we chose the composite distances as above and minimized them using again both the k-means and the pairwise merging minimization procedures. The original mixture and its reductions to 6 components using the pairwise merging and k-means algorithm is shown in Fig. 5 from which we can see that for this particular example the composite Rényi $\frac{1}{2}$-divergence minimization approach captures slightly better all the dominant modes of the mixture. The analysis of gradual reduction of the original 18-component mixture, in steps of 3 components starting with 15 and ending with 6 components, is shown in Fig. 6, where we can observe a similar situation as for the synthetic mixture, namely, calculating the ground distances between the original and reduced mixtures the pairwise merging approach showed better performance than the



(a) $\alpha = 1$

(b) $\alpha = 0.5$

Figure 5: Pairwise and k-means reduction of the people trajectory example by minimization of composite Rényi α-divergence



(a) $\alpha = 1$

(b) $\alpha = 0.5$

Figure 6: Reduction results for the people trajectory example with respect to the corresponding Rényi α-divergence for the composite distance minimization methods along with the E-M with 5000 samples

k-means. For the reference, the E-M with 5000 samples is also employed. Again all the algorithms and for each component number the reduction was repeated 50 times in order to get average performance.

6. Conclusion

In this paper we have presented a novel systematic approach to the reduction of the number of components in the mixtures of exponential families, with special emphasis on the mixtures of von Mises distributions for which explicit formulae have been presented in Section 3.2. The component number reduction problem has been formulated as an optimization problem utilizing newly proposed composite distance measures, namely the composite Rényi α-divergences, as cost functions. The benefits of using the composite Rényi α-divergences are twofold: they are computationally tractable since their value depends only on mixture parameters, and they provide an upper bound on the Rényi α-divergence itself. To solve the minimization problem we have utilized two suboptimal approaches, the generalized k-means and the pairwise merging approach. Apart from these approaches which aim to minimize the distance on the scale of the whole mixture, two local techniques were also analyzed, the joining and the West's algorithm. The presented techniques were tested and compared on a synthetic mixture and a real-world example of people trajectory shape analysis by calculating respective distances between the original and the reduced mixture and by measuring the execution time. The West's algorithm provided the most efficient approach in terms of the execution time, while the joining algorithm was more accurate in the KL distance sense. The pairwise merging scheme is a very accurate method, but computationally the most exhaustive. The k-means is computationally one of the least demanding methods, but offers variable results which depend strongly on the initial conditions. In conclusion, the results suggest that the rationale for selecting an appropriate algorithm and distance measure is an interplay between the allowable execution time, deterministic nature of the algorithm (are we willing to tolerate a variability depending on the initial conditions) and the desired accuracy of the reduction. For future work we plan to analyze such distances between mixtures of distributions to serve as an efficient criteria in measurement-to-track and track-to-track association in data association problems, and to study optimal fusion techniques of mixture distributions in state estimation problems.

Acknowledgements

This work has been supported by European Community's Seventh Framework Programme under grant agreement no. 285939 (ACROSS). We are grateful to the anonymous reviewers for useful comments and suggestions, which resulted with a significant improvement of the paper.

References

[1] G. McLachlan, D. Peel, Finite Mixture models, Wiley, 2004.

[2] S. Calderara, A. Prati, R. Cucchiara, Mixtures of von Mises distributions for people trajectory shape analysis, IEEE Transactions on Circuits and Systems for Video Technology 21 (4) (2011) 457–471.

[3] I. Marković, I. Petrović, Bearing-only tracking with a mixture of von Mises distributions, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2012, pp. 707–712.

[4] R. von Mises, Uber die 'Ganzzahligkeit' der Atomgewicht und Verwandte Fragen, Physikalische Zeitschrift 19 (1918) 490–500.

[5] K. V. Mardia, P. E. Jupp, Directional statistics, Wiley, 1999.

[6] S. R. Jammalamadaka, A. Sengupta, Topics in Circular Statistics, World Scientific, 2001.

[7] N. Fisher, Statistical Analysis of Circular Data, Cambridge University Press, 1995.

[8] N. Vasconcelos, Image indexing with mixture hierarchies, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2001, pp. 3–10.

[9] A. Nikseresht, M. Gelgon, Gossip-based computation of a Gaussian mixture model for distributed multimedia indexing, IEEE Transactions on Multimedia 10 (3) (2008) 385–392.

[10] J. Goldberger, H. Aronowitz, A distance measure between GMMs based on the unscented transform and its application to speaker recognition, in: Proceedings of Interspeech, 2005, pp. 1985–1989.

[11] D. J. Salmond, Mixture reduction algorithms for point and extended object tracking in clutter, IEEE Transactions on Aerospace and Electronic Systems 45 (2) (2009) 667–686.

[12] R. Runnalls, Kullback-Leibler Approach to Gaussian Mixture Reduction, IEEE Transactions on Aerospace and Electronic Systems 43 (3) (2007) 989–999.

[13] L.-L. S. Ong, Non-Gaussian Representations for Decentralised Bayesian Estimation, Ph.D. thesis, The University of Sydney (2007).

[14] B. Khaleghi, A. Khamis, O. K. Fakhreddine, N. R. Saiedeh, Multisensor data fusion: A review of the state-of-the-art, Information Fusion 14 (1) (2013) 28–44.

[15] J. L. Williams, P. S. Maybeck, Cost-function-based Gaussian mixture reduction for target tracking, in: Proceedings of the 6th International Conference of Information Fusion, 2003, Vol. 2, 2003, pp. 1047–1054.

[16] S. Kullback, Information Theory and Statistics, Dover Publications, Inc. New York, 1997.

[17] J. Goldberger, S. Roweis, Hierarchical clustering of a mixture model, in: NIPS, MIT Press, 2005, pp. 505–512.

[18] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, Clustering with Bregman divergences, J. Mach. Learn. Res. 6 (2005) 1705–1749.

[19] F. Nielsen, Closed-form information-theoretic divergences for statistical mixtures, in: 21st International Conference on Pattern Recognition (ICPR), 2012, pp. 1723–1726.

[20] O. Schwander, F. Nielsen, Learning mixtures by simplifying kernel density estimators, in: F. Nielsen, R. Bhatia (Eds.), Matrix Information Geometry, Springer Berlin Heidelberg, 2013, pp. 403–426.

[21] V. Garcia, F. Nielsen, Simplification and hierarchical representations of mixtures of exponential families, Signal Process. 90 (12) (2010) 3197–3212.

[22] V. Garcia, F. Nielsen, R. Nock, Levels of details for Gaussian mixture models, in: H. Zha, R.-i. Taniguchi, S. Maybank (Eds.), Computer Vision ACCV 2009, Vol. 5995 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 514–525.

[23] P. L. Dognin, J. R. Hershey, V. Goel, P. A. Olsen, Restructuring exponential family mixture models, in: INTERSPEECH'10, 2010, pp. 62–65.

[24] F. Nielsen, S. Boltz, The Burbea-Rao and Bhattacharyya centroids, IEEE Transactions on Information Theory 57 (8) (2011) 5455–5466.

[25] P. Bruneau, M. Gelgon, F. Picarougne, Parameterbased reduction of Gaussian mixture models with a variational-Bayes approach, in: 19th International Conference on Pattern Recognition, 2008.

[26] P. Bruneau, M. Gelgon, F. Picarougne, A low-cost variational-Bayes technique for merging mixtures of probabilistic principal component analyzers, Information Fusion 14 (3) (2013) 268 – 280.

[27] C. Hennig, Methods for merging Gaussian mixture components, Adv. Data Anal. Classif. 4 (1) (2010) 3–34.

[28] O. E. Barndorff-Nielsen, Information and Exponential Families in Statistical Theory, Wiley Publishers, 1978.

[29] F. Nielsen, V. Garcia, Statistical exponential families: A digest with flash cards, arXiv: 0911.4863.

[30] I. Csiszár, Why least squares and maximum entropy? an axiomatic approach to linear inverse problems, The Annals of Statistics 19 (1991) 2032–2066.

[31] S. Amari, H. Nagaoka, Methods of Information Geometry, American Mathematical Society, 2001.

[32] A. Rényi, On Measures of Entropy and Information, in: Proc. Fourth Berkeley Symp. Math. Stat. and Probability, Vol. 1, University of California Press, 1961, pp. 547–561.

[33] W. Stummer, I. Vajda, On Bregman Distances and Divergences of Probability Measures, IEEE Transactions on Information Theory 58 (3) (2012) 1277–1288.

[34] L. M. Bregman, The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming, USSR Computational Mathematics and Mathematical Physics 7 (3) (1967) 200–217.

[35] F. Nielsen, Chernoff information of exponential families, arXiv: 1102.2684.

[36] C. Villani, Optimal Transport: Old and New, Springer, 2008.

[37] N. XuanLong, Convergence of latent mixing measures in finite and infinite mixture models, The Annals of Statistics 41 (1) (2013) 370–400.

[38] S. P. Lloyd, Least squares quantization in PCM, IEEE Transactions on Information Theory 28 (2) (1982) 129–137.

[39] D. Arthur, S. Vassilvitskii, k-means++: the advantages of careful seeding, in: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, SODA '07, Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[40] D. J. Salmond, Mixture reduction algorithms for target tracking in clutter, in: Proceedings of Signal and Data Processing of Small Targets, 1990, pp. 434–445.

[41] S. Amari, α-divergence is unique, belonging to both f-divergence and Bregman divergence classes, IEEE Transactions on Information Theory 55 (11) (2009) 4925–4931.

[42] M. West, Approximating posterior distributions by mixtures, Journal of Royal Statistical Society, Series B 55 (2) (1993) 409–442.