Von Mises–Fisher Elliptical Distribution

Shengxi Li\textsuperscript{D}, Student Member, IEEE, and Danilo Mandic\textsuperscript{D}, Fellow, IEEE

Abstract—Modern probabilistic learning systems mainly assume symmetric distributions, however, real-world data typically obey skewed distributions and are thus not adequately modeled through symmetric distributions. To address this issue, a generalization of symmetric distributions called elliptical distributions are increasingly used, together with further improvements based on skewed elliptical distributions. However, existing approaches are either hard to estimate or have complicated and abstract representations. To this end, we propose a novel approach based on the von-Mises–Fisher (vMF) distribution to obtain an explicit and simple probability representation of skewed elliptical distributions. The analysis shows that this not only allows us to design and implement non-symmetric learning systems but also provides a physically meaningful and intuitive way of generalizing skewed distributions. For rigor, the proposed framework is proven to share important and desirable properties with its symmetric counterpart. The proposed vMF distribution is demonstrated to be easy to generate and stable to estimate, both theoretically and through examples.

Index Terms—Elliptical distribution, skewed distribution, von Mises–Fisher (vMF) distribution.

I. INTRODUCTION

Probabilistic distributions underpin the modeling, understanding, and prediction of a wide variety of real-world signals. Among these, the normal distribution has been a workhorse in probabilistic modeling, owing to its simple representation and mathematical tractability, while the application is justified through the central limit theorem. However, issues such as the lack of robustness and flexibility when dealing with general signals remain a serious obstacle to real-world applications. The family of elliptical distributions generalizes normal distributions and exhibits many desired properties such as simple generation, controllable robustness, and flexibility. Elliptical distributions include the normal, Cauchy, t, logistic, and Weibull distributions [1] as special cases. The well-behaved nature of elliptical distributions underpins powerful modeling tools, including those based on unimodal distributions [2], [3], mixture models [4], [5], Bayesian frameworks [6], and probabilistic graphical models [7].

Despite success, elliptical distributions inherit some of the limitations of symmetric distributions, which limits their modeling power, as in many cases such as financial, biometric, and audio scenarios, the data are not symmetric due to intrinsic coupling, systematic trend, outliers, or a small number of samples available. For example, the symmetric assumption of log returns of financial assets is fragile to financial crashes or shocks, which manifest themselves in skewed distributions [8]. We may also encounter extensive data that are intrinsically skewed in scenarios including the probability of suffering from cardiovascular diseases across ages, which exhibits a clear trend toward aged people. To address this issue, several skewed versions of the elliptical distributions have been proposed, with the majority [9]–[12] following a similar way of generalizing the normal distribution by adding a skewness weighting function [13]. Using a generalized skewed distribution does not harm the data modeling process even in the symmetric scenario, since in such cases the skewed parameter automatically vanishes in the estimation process. Although attracting attentions in recent applications [14]–[16], the type of skewed elliptical distributions results in a different stochastic representation from the (symmetric) elliptical distribution, which is prohibitive to invariance analysis, sample generation, and parameter estimation. A further extension employs a stochastic representation of elliptical distributions by assuming inner dependencies [17]; however, the added dependencies make the relationships between parameters unclear and under-determined.

In this brief, we start from the stochastic representation of elliptical distributions, and propose a novel generalization by employing the von Mises–Fisher (vMF) distribution to explicitly specify the direction and skewness, whilst maintaining the independence among the components in elliptical distributions. Such a generalization is intuitive and fully resembles the original (symmetric) elliptical distributions, which is beneficial in three aspects: 1) it admits a simple and closed-form density function, so that all the elliptical distributions can be explicitly generalized as the proposed vMF elliptical distribution; 2) it shares many desirable properties with the original elliptical distribution, including the independence between the quadratic term (or the Mahalanobis distance) and the whitened variables, the invariance property, and explicit moments; and 3) it shares the robustness properties of the elliptical distributions and can be estimated stably and efficiently, even by a naive numerical gradient descent method. This opens a new avenue for the design and implementation of robust probabilistic learning systems, such as generative models in unsupervised learning and discriminative models in supervised learning systems.

II. EXISTING GENERALIZED ELLIPTICAL DISTRIBUTIONS

A random variable, $X_e \in \mathbb{R}^n$, is said to exhibit an elliptical distribution when it has the following stochastic representation:

$$X_e \sim \mu + R \mathcal{U}$$

where $R \in \mathbb{R}$ is a non-negative scalar random variable and $\mathcal{U} \in \mathbb{S}^{n-1}$ is a random variable that is uniformly distributed on a unit sphere surface, i.e., $\mathbb{S}^{n-1} := \{x \in \mathbb{R}^n : x^T x = 1\}$. Moreover, $\mu \in \mathbb{R}^n$ and $\Lambda \in \mathbb{R}^{n \times n}$ are two constant parameters that control the distribution centers and scatter. It needs to be pointed out that the elliptical distribution is symmetric about its center, $\mu$. This is due to the fact that $R$ and $U$ in (1) are independent random variables, thus constituting a spherical distribution around $0$ via $\mathcal{R} \mathcal{U}$. The constant, $\Lambda$, transforms the sphere into an ellipse (still centered around $0$), while $\mu$ translates the center of the elliptical distribution.

When the cumulative density function (cdf) of $R$ is absolutely continuous and $\Sigma = \Lambda \Lambda^T$ is non-singular, we can write the probability density function (pdf) of the elliptical distribution as

$$p_{X_e}(x) = \det(\Sigma)^{-1/2} \cdot c_{\mu} \cdot g((x - \mu)^T \Sigma^{-1} (x - \mu))$$

where $c_{\mu} = (\Gamma(n/2))/(2^{n/2})$ is a constant solely determined by the dimension, $n$, while $g(t) = (t^2 - 1)^{-1/2}$ is a function of $t = (x - \mu)^T \Sigma^{-1} (x - \mu)$.
called the density generator, which is related to the pdf of \( R \) in (1) [11]. We denote \( X \) by \( X \sim \mathcal{E}(\mu, \Sigma, g) \). Varying \( g \) results into different types of elliptical distributions, thus achieving controllable robustness. For example, when \( g(t) \propto (1 + t)^{-((m+1)/2)} \), \( X \) becomes the Cauchy distribution, and the maximization of the log-likelihood estimation then results in a robust M-estimator against outliers [18]. The \( g(t) \) is also termed the Cauchy kernel in this case. Benefiting from its robustness to noise, the Cauchy distribution, has been applied in the matrix factorization problem for social image retrieval, as a replacement for the Gaussian assumption that is typically violated by noisy and inaccurate user tags [19]. Also in the work [20], a half-quadratic kernel Cauchy loss has been proposed and globally solved by the conjugate gradient descent method, as a way of dealing with non-Gaussian noise in nonlinear system identification.

The skewness can be modeled by adding a weighting term \( \pi(\mathbf{x} - \mathbf{\mu}) \) in (2) [9]–[11], in a way similar to the skewed normal distribution [13]. However, this type of skewness does not necessarily start from a stochastic representation, which impedes a clear interpretation of its inner relationships, generations, and moments. A further successful variant employs conditional distributions of a symmetric elliptical distribution [12], to give

\[
X_{se} \mid \lambda > 0, \quad \begin{bmatrix} \mathbf{y} \\ \lambda \end{bmatrix} \sim \mathcal{E} \left( \begin{bmatrix} \mu \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma & \beta' \\ \beta & 1 \end{bmatrix}, g \right)
\]

(3)

where the parameter \( \beta \) controls the skewness of the distribution. The form of (3) represents a typical skewed elliptical distribution, of which \( U \) has also been extended to higher dimensions [21]–[23]. Importantly, the form of (3) is invariant under quadratic forms [10] and is closed under marginalization and affine transforms; we refer to [24] for more detail. However, the estimation of the above skewed elliptical distributions can be ill-posed, especially regarding its shape (skewness) parameter. Although the singularity issue in the information matrix of shape parameter can be relieved by a centralized parametrization trick [9], the estimation of the shape parameter in this skewed version could still diverge, which calls for other penalty techniques [25]. Note that the moment estimation method employed to estimate skewed normal distributions is also inadequate for skewed elliptical distributions [26].

A further generalization that explicitly comes with the stochastic representations was proposed by Frahm [17]

\[
X_{ge} \equiv \mathbf{\mu} + \tilde{R}\mathbf{\Lambda} \mathbf{U}
\]

(4)

and has a form similar to that in (2). The difference lies in that the scalar random variable, \( \tilde{R} \), no longer needs to be non-negative and can be even negative; \( \tilde{R} \) and \( \mathbf{U} \) are also dependent, which skews the distribution. This generalization includes the skewed elliptical distribution in (3) as a special case, and is closed under affine transformation, marginalizations and even conditioning [17].

The skewness of (4) arises from the dependency between \( \tilde{R} \) and \( \mathbf{U} \). In other words, although \( \mathbf{U} \) is still distributed on a unit sphere surface, its dependency with \( \tilde{R} \) enables to vary importance along directions. Given the fact that there are only a few distributions of a scalar random variable conditioned on a vector, this flexible dependency is usually abstract and under-determined in practical modeling. More importantly, despite generalizing and skewing the elliptical distribution, this dependency also impedes the analysis of the generalized elliptical distribution, because the roles of \( \tilde{R} \), \( \mathbf{\Lambda} \) and \( \mathbf{U} \) become unclear and sometimes interchangeable. For example, in the elliptical distribution, the dispersion is uniquely dominated by \( \Sigma \), while this no longer holds in generalized elliptical distributions [17], and could lead to multiple minima/maxima when modeling data in practice.

III. GENERALIZATION VIA VON MISES-FISHER DISTRIBUTION

A. Preliminary of the vMF Distribution

Being distributed on a unit sphere surface, the vMF has been a popular choice in directional statistics [27]–[30]. The vMF distribution is determined by two parameters: \( \mu \) for the main direction and \( \tau \) for the concentration [denoted as vMF(\( \mu \), \( \tau \)). More specifically, vMF(\( \mu \), \( \tau \)) denotes a unit random variable, \( \mathbf{V} \), of the vMF distribution (i.e., \( \mathbf{V} \in \mathbb{R}^m \) and \( \mathbf{V}^T \mathbf{V} = 1 \), or equivalently \( \mathbf{V} \in \mathbb{S}^{m-1} \)) which has the following pdf:

\[
p_{\mathbf{V}}(\mathbf{v}) = \frac{\tau^{m/2-1}}{(2\pi)^{m/2} I_{(m/2)-1}(\tau)} \exp(\tau \mathbf{v}^T \mathbf{v})
\]

(5)

where \( I_{(m/2)-1}(\tau) \) represents the modified Bessel functions of the first kind. The vMF distribution, a basic element of the Stiefel manifold, has been straightforwardly applied in extensive data distribution modeling scenarios, for example, in regularizing loss distribution [31]–[33], direction-of-arrival statistics [34], beamforming statistics [35]–[37], and clustering (mixture modeling) [38]–[43]. Although achieving remarkably superior performances in task-specific applications, those works basically do not contribute with new distributions toward the vMF distribution.

Furthermore, modern statistical learning aims at increasing the power of distributions to model complicated real-world data whilst maintaining the tractability of the pdf, which simultaneously allows for straightforward and efficient estimation/optimization. This typically builds upon very carefully designed transforms from several basic elements of the vMF distribution. Several attempts to address some intrinsic limitations of the vMF distribution result in extended distributions [44]–[51]. However, all of the extended vMF variants have been designed to address directional data and therefore distribute on a sphere surface, namely, \( \mathbb{S}^{m-1} \), which is a special case of our work. To the best of our knowledge, the only exception is the Fisher–Gaussian kernel [52] that aims at accommodating curvature information into distributions. Including the vMF distribution, the proposed Fisher–Gaussian kernel consists of three random variables and then linearly adds them up to form a new distribution. However, the sole summation operation in the Fisher–Gaussian kernel limits the modeling efficiency and capacity in applications. Furthermore, the Fisher–Gaussian kernel is basically different from the elliptical distribution, so that it loses connections to many important distributions including Student-t, Cauchy, and \( \alpha \)-stable distributions.

To the best of our knowledge, our work sets out the first broad class of skewed elliptical distributions that preserves all fundamental and desirable properties of the symmetric elliptical distribution, and at the same time, it generalizes well to advanced distributions, including both directional and non-directional, as well as both symmetric and nonsymmetric distributions. As further elaborated shortly, in addition to being well-behaved, powerful, and general in data modeling, the derivation of our work is extremely intuitive, which makes it possible to inherently enjoy many desirable properties.

B. vMF Elliptical Distribution

Being distributed on a unit sphere surface, the vMF is a popular choice in directional statistics [27]–[30]. The vMF distribution is determined by two parameters: \( \mu \), for the main direction and \( \tau \) for the concentration [denoted as vMF(\( \mu \), \( \tau \)). Therefore, it is natural and beneficial to replace \( \mathbf{U} \) in (2) by the vMF distribution as a way of explicitly expressing the direction information. We thus propose a new type of generalization on the elliptical distributions in the form

\[
X = \mathbf{\mu} + \tilde{R}\mathbf{\Lambda} \mathbf{V}
\]

(6)
where $\mathbf{V}$ denotes a random variable satisfying the vMF distribution $\text{vMF}(\mu_*, \tau)$. In our definition, $\mathbf{R}$ is the same as that in (2), i.e., non-negative and independent of $\mathbf{V}$. More importantly, when $\tau \to 0$, the vMF distribution approaches the uniform distribution on a unit sphere $\mathbf{U}$, and consequently (6) degenerates into the symmetric elliptical distribution. This generalization maximally preserves the formats and desirable properties of the symmetric elliptical distribution, such as the independence and clear physical meaning of each part. In other words, in our vMF elliptical distribution, $\mu$ closely relates to the data location, $\mathbf{R}$ governs the tails and $\mathbf{V}$ the directions (skewness). As shall be shown shortly, this is beneficial in both theoretical analysis and practical estimator settings.

The pdf of $\mathbf{X}$ in (6) can be obtained in a closed-form as

$$p_{\mathbf{X}}(x) = \det(\Sigma)^{-1/2} \cdot p_{\mathbf{V}}\left(\frac{\Sigma^{-1/2}(x - \mu)}{\sqrt{t}}\right) \cdot g(t)$$

(7)

where $t$ represents the Mahalanobis distance i.e., $t = (x - \mu)^T \Sigma^{-1}(x - \mu)$ and $p_{\mathbf{V}}(\cdot)$ is the pdf of vMF distribution $\text{vMF}(\mu_*, \tau)$. We provide the proof of (7) in Appendix A. An intuitive way of understanding $\mathbf{X}$ is the skewness part of our vMF elliptical distribution. We first address in Proposition 1 the invariance of location, $\mathbf{R}$ controls the dispersion, $\mathbf{V}$ the directions (skewness). As shall be shown shortly, this is beneficial in both theoretical analysis and practical estimator settings.

**Proposition 1:** The skewness of the distribution is invariant to the quadratic term $(\mathbf{X} - \mu)^T \Sigma^{-1}(\mathbf{X} - \mu)$.

*Proof:* From (6), we have

$$(\mathbf{X} - \mu)^T \Sigma^{-1}(\mathbf{X} - \mu) = \mathcal{R}^2 \mathcal{V}^T \mathbf{V} = \mathcal{R}^2.$$

(9)

Because $\mathcal{R}$ is independent of $\mathcal{V}$, the quadratic term is irrelevant to the skewness part of our vMF elliptical distribution.

This completes the proof.

The property in Proposition 1 is the basis for satisfying the distributional invariance [53], which provides convenience in applications such as hypothesis testing and dealing with sampling bias [10].

Given the fact that $(\Sigma^{-1/2}(\mathbf{X} - \mu))/(\mathbf{X} - \mu)^T \Sigma^{-1}(\mathbf{X} - \mu)) \equiv \mathcal{V}$, from (6) we can easily obtain the following result based on the proof of the Proposition 1.

**Proposition 2:** The quadratic term $(\mathbf{X} - \mu)^T \Sigma^{-1}(\mathbf{X} - \mu)$ is independent of the whitened random variable $((\Sigma^{-1/2}(\mathbf{X} - \mu))/(\mathbf{X} - \mu)^T \Sigma^{-1}(\mathbf{X} - \mu))$.

This independence property plays a crucial role in the calculation of robustness and other related proofs. Furthermore, due to the neat generalization and clear physical meaning of our vMF elliptical distribution, we can write the moments in a closed-form, with the first two moments are addressed in Proposition 3.

**Proposition 3:** Given the generalization of (6), we have

$$E[\mathbf{X}] = \mu + (\rho_0(\tau)E[\mathbf{R}]A)\mu_a$$

$$\text{Var}[\mathbf{X}] = E[\mathbf{R}]^2 \rho_0(\tau) \Sigma$$

$$- E[\mathbf{R}]^2 (1 - \frac{m}{\tau} \rho_0(\tau) - \rho_0^2(\tau)) A \mu_a \mu_a^T A^T$$

(10)

where $\rho_0(\tau) \in (0, 1)$ is the ratio of two modified Bessel functions of the first kind, given by

$$\rho_0(\tau) = \frac{I_{m/2}(\tau)}{I_{m/2-1}(\tau)}.$$  

(11)

*Proof:* Given the stochastic representation of (6), we have

$$E[\mathbf{X}] = \mu + E[\mathbf{R}]A\mathcal{V}.$$  

(12)

More importantly, due to the independence of $\mathbf{R}$ and $\mathbf{V}$, we further have

$$E[\mathbf{X}] = \mu + E[\mathbf{R}]A E[\mathbf{V}] = \mu.$$  

(13)

Recall that $\mathbf{V}$ is the vMF parametrized as $\text{vMF}(\mu_*, \tau)$. Its first-order moment is calculated by $E[\mathbf{V}] = \rho_0(\tau)\mu_*$ [54]. Thus, the first-order moment of the vMF elliptical distribution can be obtained as

$$E[\mathbf{X}] = \mu + (\rho_0(\tau)E[\mathbf{R}]A)\mu_a.$$  

(14)

Similarly, the second-order moment of the vMF distribution is given by $E[\mathbf{V}\mathbf{V}^T] = (\rho_0(\tau)/\tau) I + (1 - (m/\tau)\rho_0(\tau))\mu_a^T \mu_a$ [54], so that we arrive at

$$\text{Var}[\mathbf{X}] = E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}]^T$$

$$= E[\mu\mu^T + \mu^T R \mathcal{V}^T \mathcal{V} \mu + R^2 \mathcal{V}^T \mathcal{V}]$$

$$- (\mu\mu^T + E[R]E[\mathbf{V}]E[\mathbf{V}^T] + E[R]AE[\mathbf{V}]E[\mathbf{V}])$$

$$= E[R^2] E[\mathbf{V}] E[\mathbf{V}]^T - E[\mathbf{V}] E[\mathbf{V}]^T$$

$$= E[R^2] \rho_0^2(\tau) \Sigma$$

$$- E[R^2] (1 - \frac{m}{\tau} \rho_0(\tau) - \rho_0^2(\tau)) A \mu_a \mu_a^T A^T.$$  

(15)

This completes the proof.

Furthermore, when $\tau \to 0$, the following bounds hold, according to the Taylor series of the modified Bessel function of the first kind $(I_{m/2}(\tau) \approx (1/(\Gamma(m/2 + 1)))(\tau^{m/2}))$ for $b \to 0$:

$$\lim_{\tau \to 0} \rho_0(\tau) = \frac{I_{m/2}(\tau)}{I_{m/2-1}(\tau)} = \frac{\Gamma(m/2)}{2(m/2)\Gamma(m/2)} = 1.$$  

$$\lim_{\tau \to 0} \rho_0(\tau) = \frac{\Gamma(m/2)}{2(m/2)\Gamma(m/2)} = \frac{1}{m}.$$  

(16)

Therefore, from Proposition 3, when $\tau \to 0$, we have $E[\mathbf{X}] = \mu$ and $\text{Var}[\mathbf{X}] = (E[R^2])/m \Sigma$, which are exactly the moments of the symmetric elliptical distributions [17].

**D. Optimization**

Given the independent and identically distributed (i.i.d.) nature of the samples $[\mathbf{x}_i]_{i=1}^n$, our vMF elliptical distribution can be estimated by maximizing the log-likelihood (MLL), which is equivalent to minimizing the Kullback–Leibler divergence between the parametric and empirical distributions. It also needs to be pointed out that other distance measures can be employed in estimation, such as the Wasserstein distance, a subject of future work.

For notational simplicity, we shall denote the whitened term as $z = ((\Sigma^{-1/2}(\mathbf{x} - \mu))/\sqrt{t})$ and the quadratic term as
$t = (x - \mu)^T \Sigma^{-1} (x - \mu)$. We obtain the explicit pdf of our vMF elliptical distribution by combining (5) and (7) as

$$p_X(x) = \frac{\text{det}(\Sigma)^{-1/2}}{2\pi^{m/2} I_{(m/2)-1}(\tau)} \exp(\tau \mu^T z_i) \cdot g(t).$$

Then, due to the simple form of the pdf of our generalized distribution, we can omit the constant term and write the log-likelihood from (7) as

$$L(\mu, \Sigma, \mu_*, \tau) = \sum_{i=1}^{n} \log p_X(x_i)$$

$$= \sum_{i=1}^{n} \log \left( \frac{\text{det}(\Sigma)^{-1/2}}{2\pi^{m/2} I_{(m/2)-1}(\tau)} \exp(\tau \mu^T z_i) \cdot g(t_i) \right)$$

$$\propto -\frac{n}{2} \ln \text{det}(\Sigma) + n \left( \frac{m}{2} - 1 \right) \ln \tau$$

$$-n \ln I_{(m/2)-1}(\tau) + \sum_{i=1}^{n} \tau \mu^T z_i + \sum_{i=1}^{n} \ln g(t_i).$$

Furthermore, as $||\mu_*|| = 1$, this leads to constrained optimization, whereby we reparametrize (18) by $\nu = \tau \mu_*$ (where $\tau = ||\nu||_2$ and $\mu_* = (\nu/||\nu||_2)$) as

$$L(\mu, \Sigma, \nu) \propto -\frac{n}{2} \ln \text{det}(\Sigma) + n \left( \frac{m}{2} - 1 \right) \ln ||\nu||_2$$

$$-n \ln I_{(m/2)-1}(||\nu||_2) + \sum_{i=1}^{n} \nu^T z_i + \sum_{i=1}^{n} \ln g(t_i).$$

By taking the derivatives with respect to $\mu$, $\Sigma$, and $\nu$, and setting them to $0$, we obtain the following optimum, with the detailed proof given in Appendix B:

$$\mu^* = \frac{\sum_{i=1}^{n} \left( \left( v_i \nu - 2 \psi(t_i) \right) x_i - \frac{1}{\tau} \Sigma \nu^T v \right)}{\sum_{i=1}^{n} \left( v_i \nu - 2 \psi(t_i) \right)^2 - 2 \psi(t_i)}$$

$$\Sigma^* = \frac{\sum_{i=1}^{n} \left( v_i \nu - 2 \psi(t_i) \right) (x_i - \mu)(x_i - \mu)^T - \Sigma \nu^T v \nu^T}{n}$$

$$\nu^* = \rho_{\nu}(\cdot) \left( \frac{1}{n} \sum_{i=1}^{n} z_i \right) \left( \frac{1}{||\sum_{i=1}^{n} z_i||_2} \right).$$

In (20), $\rho_{\nu}(\cdot)$ is an inverse function of $\rho_\mu(\cdot)$ and $\psi(t) = ((\partial g(t)/\partial t)/g(t))$. Moreover, when $\tau \rightarrow 0$, $\nu^* \rightarrow 0$, and the optimal $\mu$ and $\Sigma$ also represent the optimum of the symmetric elliptical distribution.

### IV. Numerical Results

#### A. Synthetic Datasets

In our experiments on synthetic datasets, we tested five degrees of concentration, i.e., $\tau = \{2, 4, 6, 8, 10\}$, together with five different dimensions ($m = \{2, 4, 8, 16, 32\}$), under two types of distributions, i.e., vMF Gaussian and vMF Cauchy distributions, thus having $5 \times 5 = 50$ test cases. For each test case, we generated 1000 synthetic samples from distributions of randomly chosen parameters $(\mu, \mu_*, \Sigma)$. The only controlled parameter was the eccentricity of the distribution, which is the ratio of the maximum eigenvalue over the minimum eigenvalue of $\Sigma$; we set it smaller than 4 to avoid the majority of samples being aligned on low-dimensional spaces. For the synthetic data, in each test case, we performed estimation over ten trials, with random initializations. Standard steepest gradient descent was used to update $\mu$ and $\Sigma$, with the learning rate of 0.01, while $\mu_*$ and $\tau$ were optimized by a truncated Newton method proposed in [30]. The mean values of likelihood error ratios are plotted in Fig. 1, with the likelihood error ratio defined as $r = \left( (l_{\text{est}} - l_{\text{true}})/l_{\text{true}} \right)$, while $l_{\text{est}}$ denotes the log-likelihood given by the estimated parameters and $l_{\text{true}}$ for the true log-likelihood.

From Fig. 1, we observe that different from other generalized elliptical distributions, generating our vMF elliptical distributions is straightforward to implement and easy to control. More importantly, by virtue of its simple estimation method, the error ratios of estimating both vMF Gaussian and Cauchy distributions were both lower than 0.07, and even lower than 0.03 except for the case $m = 2$. However, the obtained small error ratios also verify that the proposed vMF elliptical distribution is not only easy to generate but is also stable. Moreover, Fig. 1 also shows that increasing either the dimensions $m$ or the degrees of concentration $\tau$ did not enlarge the estimated error, thus validating the consistency of our simple yet robust and effective estimation. Employing advanced numerical solvers such as conjugate gradient descent and trust region method could further improve the estimation, a subject of future work.

#### B. Real-World Datasets

For comprehensive evaluations in real-world applications, We have now extensively compared our method against additional three state-of-the-art baselines on additional five real-world applications/datasets that are associated with health, chemistry, crimes, plants, and finance scenarios. We have also further evaluated a broader class of vMF elliptical distributions, by additionally comparing the vMF-$\mu$, vMF-Laplace, and vMF-Logistic distributions, to further highlight the
The power of our proposed framework regarding both flexibility and efficient estimation. Generally speaking, we first evaluated our work on Australian Institute of Sport (AIS) and Piedmont wines datasets, two standard datasets within the skewed normal (SN) package in R [55], together with iris [56] and USA Crimes [57] datasets that are two widely employed real-world datasets in multivariate distribution modeling. For large-scale applications, we further evaluated our method over stocks of 17 companies from the SPX500 index, ranging across 20 years.

As for comparing methods, we found that there are few comparisons across differently proposed multivariate skewed distributions in the literature, the fact also noticed in [24], and new proposals [12], [13], [17], [22] are mostly presented without estimation strategies. Although using smart search methods, such as simplex search [58], might find a possible solution when estimating parameters, this may be problematic in high dimensions, especially when estimating the positive semi-definite covariance matrix. In contrast, our method enjoys much better efficiency, stability, and convergence behavior, since it explicitly calculates the gradients toward the optimum. To the best of our knowledge, the SN package is the only publicly available implementation that provides estimation toward the optimum. To the best of our knowledge, the SN package is the only publicly available implementation that provides estimation strategies. Although using smart search methods, such as simplex search [58], might find a possible solution when estimating parameters, this may be problematic in high dimensions, especially when estimating the positive semi-definite covariance matrix.

Toward evaluations, we have now reported the final log-likelihood $l_{\text{est}}$ as the main metric because this is equivalent to measuring the Kullback–Leibler divergence between the estimated and the ground truth data distributions, the ultimate objective in density estimation [61]. More importantly, over the same dataset, an optimized distribution with the maximized log-likelihood is of significant potential to extensive task-specific applications. For example, a generalized/skewed distribution, equipped with the superior estimation, can remarkably improve the regression models, where the estimated location is tightly related to the predictions [24, Sec. 4.1]. The covariance, obtained from a maximized log-likelihood, is also crucial in various applications including finance risk management, portfolio optimization, etc., [24, Sec. 5.2]. Since our work aims at proposing a broad class of well-behaved distributions, given limited space of brief, we focus on evaluating the very fundamental task/application, namely, the ability of fitting real-world data, and leave further investigations over specific applications as our future work.

More specifically, when comparing over the AIS, Piedmont wines and iris datasets, we remove the attributes of discrete levels for estimating continuous distributions. The removed attributes are the [sex and sport] items in the AIS dataset, the [wine name] attribute in the Piedmont wines dataset and the species in the iris dataset. The resulting data matrices ($n \times m$) for the AIS, Piedmont wines, iris, and USA Crimes datasets are 202 × 11, 178 × 27, 150 × 4 and 50 × 4. The estimated $l_{\text{est}}$ of each method is reported in Table I. These results establish that, benefiting from the intuitive construction, our method achieves the highest log-likelihood against the existing skewed distributions, thus verifying the best accommodation across all the real-world datasets. Another advantage is that besides the skewed Gaussian, $t$ and Cauchy distributions, our vMF elliptical distribution is flexible, so that it allows to straightforwardly obtain various other skewed distributions that are guaranteed by stable and efficient estimation.

Furthermore, we have evaluated our method on the SPX500 financial stock data, a relatively large-scale application. We here randomly selected 17 companies, and together with the return, constituted a dimension of 18.1 Moreover, the stock prices range from May 8, 2000, to June 5, 2019, and therefore constitute the data matrix of dimension 4799 × 18 by removing missing dates (for example, the weekend). For a comprehensive comparison, besides the final estimated $l_{\text{est}}$, we also fit the data that consist of paired companies, as illustrated in Fig. 2. From this figure, observe that our method achieved the highest $l_{\text{est}}$ against the overall and paired companies, compared to the existing state-of-the-art skewed distributions. By inspection, we can also find that the vMF-$t = 10$ distribution performs much better than the vMF-Gaussian distribution, a similar trend that also appears on skewed normal against skewed $t = 10$.

1Those 18 dimensions index the cost return, BWA, MCD, PPG, PPL, OXY, IFF, MAS, CCL, AAPL, CAG, NEM, NOV, BAX, CI, EFX, SYY, and CVS companies, respectively.

| Table I | Estimated Average Log-Likelihood $l_{\text{est}}$ of Multivariate Skewed Distributions Against the AIS, Piedmont Wines (P. Wines for Short), Iris, and USA Crimes Datasets |
|---------|-------------------------------------------------------------------------------------------------|
|          | AIS | P. wines | iris | USA Crimes |
| Skewed normal [13], [25] | 8.83 | 1.94 | 0.06 | -1.37 |
| Skewed $t = 10$ [59] | 8.84 | 2.05 | 0.01 | -1.45 |
| Skewed Cauchy [60] | 8.22 | 1.32 | -0.37 | -1.73 |
| vMF-Gaussian | 15.29 | 31.71 | 0.58 | -0.93 |
| vMF-$t = 10$ | 15.06 | 32.00 | 0.53 | -0.95 |
| vMF-Cauchy | 13.70 | 31.23 | 0.02 | -1.31 |
| vMF-Laplace | 14.00 | 31.50 | 0.23 | -1.08 |
| vMF-Logistic | 20.94 | 47.04 | 2.50 | 0.99 |

![Fig. 2. Estimated log-likelihood $l_{\text{est}}$ between paired companies on the SPX500 financial stock data. Note that the top-right elements denote the estimation of the skewed normal [13], [25] in (a) and skewed $t = 10$ [60] in (b), whereas the bottom-left elements are from our vMF-Gaussian and vMF-$t = 10$ distributions. The overall $l_{\text{est}}$ of the skewed normal [13], [25] and skewed $t = 10$ [60] are 14.87 and 17.14, whilst those for vMF-Gaussian and vMF-$t = 10$ are 32.24 and 33.29, respectively.](image-url)
distributions. This might be due to the fact that the companies have experienced for example, 2008 financial crash, scenarios where the heavily tailed distribution can well accommodate the impulsive behaviors. Our work, being able to further include other flexible distributions, is therefore also applicable to the related fields of statistical signal processing and applications.

V. Conclusion

We have proposed a novel generalized elliptical distribution, termed vMF elliptical distribution, in order to generalize symmetric elliptical distributions and equip them with physical meaning in the analysis, simple representations, and the ability to represent general skewed distributions. This has been achieved in an intuitive manner, starting from the stochastic representation, which has enabled the proposed vMF elliptical distribution to exhibit many desirable properties, together with an explicit pdf. We have also introduced optimal parameter estimation into the vMF elliptical distributions, and have experimentally validated that even a basic steepest gradient descent method can achieve consistent and stable low estimation errors. This opens new avenues for both future applications and research on skewed distributions, especially in the context of probabilistic learning machines.

APPENDIX

A. Proof of (7)

Note that \( R \) is independent of \( \mathbf{V} \), so that the following holds:

\[
p_{\mathcal{R}, \mathbf{V}}(r, v) = p_{\mathcal{R}}(r) \cdot p_{\mathbf{V}}(v).
\]

We next define a random variable by a multiplication \( \mathbf{Y} = f(\mathcal{R}, \mathbf{V}) = \mathcal{R} \cdot \mathbf{V} \). It needs to be pointed out that \( \mathbf{V} \) is a unit random vector, which means that the multiplication function \( f \) is an injection due to \( |\mathcal{R}, \mathbf{V}| = |\mathcal{R} \cdot \mathbf{V}| \). Given the change of variables formula [62], we have that the matrix volume is \( |\mathbf{Y}|_2^{-(n-1)} \). Thus, the pdf of \( \mathbf{Y} \) can be obtained as

\[
p_{\mathbf{Y}}(y) = p_{\mathcal{R}, \mathbf{V}}(|\mathbf{y}|_2, \frac{y}{|\mathbf{y}|_2}) \cdot |\mathbf{y}|_2^{-(n-1)} = p_{\mathcal{R}}(|\mathbf{y}|_2) \cdot p_{\mathbf{V}}(y) \cdot |\mathbf{y}|_2^{-(n-1)}.
\]

By employing the linear transform, \( \mathbf{X} = \mathbf{Y} = \Sigma^{1/2} \mathbf{Y} \), the pdf of our vMF elliptical distribution can be obtained by substituting \( t = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) = |\mathbf{y}|_2^2 \), as

\[
p_{\mathbf{X}}(x) = \det(\Sigma)^{-\frac{1}{2}} \cdot |\mathbf{y}|_2^{-n} \cdot p_{\mathbf{V}}\left(\frac{\Sigma^{-1/2}(\mathbf{x} - \mu)}{\sqrt{\mathbf{v}}}ight) \cdot f^v(g(\mathbf{t})).
\]

B. Proof of (20)

By using the relationship of the modified Bessel function of the first kind \( (\partial I_{\nu}(b))/\partial b = I_{\nu+1}(b) + (a/b)I_{\nu}(b) \), we can simplify the derivative of \( \nu \) as

\[
\frac{1}{n} \frac{\partial L(\mathbf{\mu}, \Sigma, \mathbf{v})}{\partial \mathbf{v}} = \left( \frac{m}{2} - 1 \right) \mathbf{v} \left( \frac{|\mathbf{v}|}{|\mathbf{v}|} \right)^2 + \sum_{i=1}^{n} z_i \\
\begin{align*}
&= \frac{1}{n} \left( \frac{m}{2} - 1 \right) \mathbf{v} \left( \frac{|\mathbf{v}|}{|\mathbf{v}|} \right)^2 \\
&- \frac{I_{\nu+1}(|\mathbf{v}|)}{|\mathbf{v}|} - \frac{I_{\nu}(|\mathbf{v}|^2)}{2} \left( \frac{m}{2} - 1 \right) I_{\nu+1}(|\mathbf{v}|) \\
&= -\frac{1}{n} \left( \frac{m}{2} - 1 \right) \mathbf{v} \left( \frac{|\mathbf{v}|}{|\mathbf{v}|} \right)^2 + \sum_{i=1}^{n} z_i.
\end{align*}
\]

By setting the derivative in (24) to 0, we arrive at the optimal \( \mathbf{v} \) in the form

\[
\mathbf{v}^* = \left( \frac{1}{n} \sum_{i=1}^{n} z_i \right) \mu - \frac{1}{n} \sum_{i=1}^{n} z_i \mu - \frac{1}{n} \sum_{i=1}^{n} z_i \mu.
\]

The optimal \( \mu \) can be calculated in a similar way

\[
0 = \frac{\partial L(\mathbf{\mu}, \Sigma, \mathbf{v})}{\partial \mathbf{\mu}} = \sum_{i=1}^{n} \left( -\Sigma^{-1} \Sigma^{-1}(\mathbf{x}_i - \mu) \right) - 2\psi(t_i) \Sigma^{-1}(\mathbf{x}_i - \mu).
\]

with the optimum at

\[
\mu^* = \sum_{i=1}^{n} \left( \frac{\xi_i^2}{\Sigma} - 2\psi(t_i) \right) / \sum_{i=1}^{n} \xi_i^2.
\]

To find the optimum of \( \Sigma \), it is convenient to decompose \( \Sigma \) into \( \Lambda \Lambda^T \) by the Cholesky decomposition because the square root of the positive-definite matrix \( \Sigma \) is guaranteed. Then, by taking the derivatives with regard to \( \Lambda \), we have

\[
0 = \frac{\partial L(\mathbf{\mu}, \Sigma, \mathbf{v})}{\partial \Lambda} = -n \Lambda^T + \sum_{i=1}^{n} \left( \frac{1}{t_i} \xi_i^2 \Sigma^{-1}(\mathbf{x}_i - \mu)^T \Lambda^{-T} - \xi_i \psi(t_i)^2 \Lambda^{-T} \right)
\]

\[
- \sum_{i=1}^{n} \left( 2\psi(t_i) \Sigma^{-1}(\mathbf{x}_i - \mu)^T \Lambda^{-T} \right).
\]

After some simplifications, the optimal \( \Sigma \) can be easily obtained in a closed-form as

\[
\Sigma^* = \sum_{i=1}^{n} \left( \frac{\xi_i^2}{\Sigma} - 2\psi(t_i) \right) (\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)^T - \sum_{i=1}^{n} \xi_i \psi(t_i)^2
\]

\[
\frac{n}{n}.
\]

REFERENCES

[1] K. W. Feng, S. Kotz, and K. W. Ng, Symmetric Multivariate and Related Distributions. London, U.K.: Chapman & Hall, 1990.

[2] A. Renaux, P. Forster, E. Chaumette, and P. Larzabal, “On the high-SNR conditional maximum-likelihood estimator of unimodal elliptical distribution,” IEEE Trans. Signal Process., vol. 54, no. 12, pp. 4840–4843, Dec. 2006.

[3] S. Zozor, C. Ren, and A. Renaux, “On the maximum likelihood estimator for unimodal elliptical distributions in the high signal-to-noise ratio regime,” IEEE Signal Process. Lett., vol. 25, no. 6, pp. 883–887, Jun. 2018.

[4] S. Li, Z. Yu, M. Xiang, and D. P. Mandic, “Solving general elliptical mixture models through an approximate Wasserstein manifold,” in Proc. AAAI, 2020, pp. 4658–4666.

[5] S. Li, Z. Yu, and D. Mandic, “A universal framework for learning the elliptical mixture model,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 7, pp. 3181–3195, Jul. 2021.

[6] S. Zozor and C. Vignat, “On the MAP estimation in the context of elliptical distributions,” in Proc. 17th Eur. Signal Process. Conf.

[7] L. Stanković et al., “Data analytics on graphs Part III: Machine learning on graphs,” from graph topology to applications,” Found. Trends Mach. Learn., vol. 13, no. 4, pp. 332–530, 2020.

[8] C. J. Adcock, “Asset pricing and portfolio selection based on the multivariate skew-normal distribution,” J. Roy. Stat. Soc., B, Stat. Methodol., vol. 61, no. 3, pp. 579–602, Aug. 1999.
