2024 American Control Conference (ACC)
July 8-12, 2024. Toronto, Canada

Closed-Form Information-Theoretic Roughness Measures for Mixture Densities

Uwe D. Hanebeck, Daniel Frisch, and Dominik Prossel

***Abstract*— In estimation, control, and machine learning under uncertainties, latent variables are usually described by a probability density function (pdf). The optimal reconstruction of a continuous pdf from given samples or moments is an important and ubiquitous task. Unfortunately, it typically results in an underdetermined optimization problem, as the pdf is not fully constrained by the given samples or moments. For regularization, we use Fisher Information (FI) that acts as a roughness measure, i.e., selects the smoothest pdf fulfilling the constraints, in an information-theoretic sense. For the important class of mixture densities, FI can only be computed numerically. In this paper, we derive a closed-form solution for FI for mixtures by transforming the problem to the space $\mathcal{R}$ of root mixtures (RMs). This results in a tandem processing scheme simultaneously working in the original mixture space $\mathcal{M}$ and the corresponding RM space: The density parameters are optimized in root mixture space based on the closed-form FI. The desired constraints are evaluated in the original mixture space $\mathcal{M}$.**

I. Introduction

In many applications, we face the problem of reconstructing the full pdf $\tilde{f}(\underline{x})$ describing a vector $\underline{\boldsymbol{x}}$ of desired latent variables from a set of constraints or specifications on the pdf. Application include (1) replacing a moment representation of $\tilde{f}(\underline{x})$ with a continuous density representation $f(\underline{x})$, (2) estimating a continuous pdf representing a given set of samples, (3) increasing the number of parameters of density representation, e.g., increasing the number of components of a mixture pdf, and (4) interpolating a density from given density values. Common to these examples is the incomplete specification of the desired pdf, which leads to an underdetermined optimization problem.

In general, the specifications can simply be lower-order moments, e.g., mean, covariance, and samples of $\tilde{f}$. However, they may also include higher-order moments, density values $\tilde{f}(\underline{x}_i)$ and/or its derivatives at selected $\underline{x}_i$, and probabilities of $\underline{\boldsymbol{x}}$ being in certain regions.

Given the specifications on the pdf $\tilde{f}(\underline{x})$, we want to reconstruct an approximation $f(\underline{x})$ of $\tilde{f}(\underline{x})$ without artificially adding unwarranted information[1]. In summary, we desire the least-informative pdf given the specification.

II. Problem Formulation

First, we are given the set specifications $S_i(f) = 0\,,\ S_i \in \mathcal{S}$[2] about the desired pdf $f(\underline{x})$. We assume an underlying pdf $\tilde{f}(\underline{x})$ that is unknown and that we would like to approximate with the desired pdf $f(\underline{x})$[3].

Uwe D. Hanebeck, Daniel Frisch, and Dominik Prossel are with the Intelligent Sensor-Actuator-Systems Laboratory (ISAS), Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology (KIT), Germany. Please send correspondence to Uwe.Hanebeck@kit.edu.

Second, we are given a certain structure of the pdf $f(\underline{x})$. This can be achieved by confining $f(\underline{x})$ to a certain class of densities, say, spherically invariant densities [25]. More concrete, the pdf could be fixed to being of a specific type, say, a Gaussian density. Here, our focus is on pdfs from the class of mixture densities (MDs) with a given number L of components

$$f(\underline{x}) = \sum_{i=1}^{L} w_i f_i(\underline{x}) \tag{1}$$

with positive weights $w_i > 0$, $i \in 1{:}L$, summing up to one

$$\sum_{i=1}^{L} w_i = 1\ , \tag{2}$$

and mixture component pdfs $f_i(\cdot)$, $i \in 1{:}L$. As a concrete density type, we will consider Gaussian components.

We want to find a MD $f(\underline{x})$ that fulfills the given specifications. Specifically we desire appropriate mixture weights $w_i > 0$, $i \in 1{:}L$, and parameters of the component pdfs $f_i(\cdot)$, $i \in 1{:}L$.

When the number of parameters is larger than the number of specifications, which is often the case, the desired MD $f(\underline{x})$ is not uniquely specified. We propose to select the MD with the lowest information content. The FI will be used for that purpose, which will be shown to act as a roughness measure, i.e., favors smooth pdfs.

III. State-of-the-Art

In this section, we briefly review the state-of-the-art in calculating least-informative pdfs. Given a set of specifications on the pdf, a least-informative pdf does not add more information than necessary. A pdf with little information content intuitively exhibits a high level of smoothness or equivalently a low level of roughness. We will focus on methods based on curvature, entropy, and FI as a basis for the derivation of the proposed new method in Sec. IV.

A. Mean Curvature

At first look, mean curvature would be a good candidate for quantifying smoothness. The problem is that integrating over the curvature is numerically difficult. Hence, curvature is often approximated by the second derivative of the considered pdf. Expressions of this type are used as roughness penalties for smoothing splines [26, p. 177].

For simplicity, we only consider the scalar case, which is already quite complex. The local curvature of f is defined

as[4]

$$k\big(f(x)\big) = \frac{f''(x)}{\Big(1 + \big(f'(x)\big)^2\Big)^{3/2}} \ . \tag{3}$$

The mean squared curvature is given as the integral over the local curvature as

$$k(f) = \int_{\mathbb{R}} k^2\big(f(x)\big)\,\mathrm{d}x \ , \tag{4}$$

which cannot be solved in closed form for densities of interest here. For this reason, the local curvature is often approximated with the second derivative of f as $k(f(x)) = f''(x)$. In higher dimensions, this can be generalized to $k(f(\underline{x})) = \nabla^2 f(\underline{x})$ with Hessian operator $\nabla^2 = \nabla \cdot \nabla^\top$. The resulting mean squared curvature

$$k(f) = \int_{\mathbb{R}^D} \left|\nabla^2 f(\underline{x})\right|^2 \mathrm{d}\underline{x} \tag{5}$$

can often be calculated in closed form, e.g., for Gaussian or Gaussian Mixture (GM) densities $f(\underline{x})$. However, although this simplified mean curvature might be useful for, e.g., spline smoothing, it is not a good indicator of smoothness for pdfs. The reason is that in a typical reference situation, i.e., minimization of the roughness of a density with zero mean and unit variance, minimization of the curvature does not yield a Gaussian density as would be expected. On the other hand, maximization of the entropy or minimization of the FI yields the expected Gaussian density in this case as we will see in Subsec. III-B and Subsec. III-C.

Summary: The mean curvature of a general density according to (3) and (4) is difficult to compute even in the univariate case. This is exacerbated for mixture densities. For this reason, the curvature is often approximated by the second derivative of $f(x)$ in 1D. In the multivariate case, the definition of curvature is even more complicated, which is another reason for approximation by the Hessian of $f(\underline{x})$. Unfortunately, the simplified mean curvature, although easily computable for mixtures, does not yield the expected results in simple reference cases.

B. Entropy

Most methods for finding the least-informative pdf under given specifications use the principle of maximum entropy. It was conceived by Jaynes in 1957 [13, 14]. It received its share of criticism [7], but was eventually adopted by the research community. One example is the use of maximum entropy for the trigonometric moment problem and orthogonal polynomials [15].

For a continuous density $f(\underline{x})$, the so-called differential entropy is defined as

$$\mathrm{E}\{-\log\left(f(\underline{x})\right)\} = -\int_{\underline{x}\in\mathbb{R}^D} f(\underline{x}) \log\left(f(\underline{x})\right) \mathrm{d}\underline{x} \ . \tag{6}$$

The differential entropy has many nice properties. However, it is an ad-hoc generalization of the famous Shannon entropy for discrete random variables to the continuous setting. It can assume negative values and is not invariant under a change of variables.

A table with entropy expressions for standard pdfs is provided in [16]. Expressions for the entropy in multivariate settings are given in [6], but this does not include mixture distributions. Because of the logarithm in the expression of the differential entropy used in maximum entropy methods, computation is complicated beyond simple cases such as Gaussian densities. This is especially a problem for mixture densities as these lead to logarithms of sums. As a result, entropy is often calculated via numerical integration, e.g., Monte Carlo, which leads to a high computational load in multivariate settings.

This led to an extensive development of approximations and bounds, with a strong focus on mixture densities. Differential entropy for GMs is approximated in [11] by a Taylor-series expansion of the logarithm of the GM around each component mean. For large component variances, splitting of components is required to maintain a desired accuracy. In [10], a deterministic sample representation is approximated by a piecewise constant density to facilitate the computation of the relative entropy. A piecewise constant approximation of a GM is proposed in [29]. Sharp bounds on the so-called entropy concavity deficit are derived in [18], i.e., the difference between the mixture entropy and the sum of component entropies. In [20], lower and upper bounds on the differential entropy for the special case of GMs are provided when all components have identical variances and only differ in their means and weights. For a symmetric GM with two components with equal weights and equal variances, lower and upper bounds on the differential entropy are given in [19].

Summary: As it requires integration over a logarithm of $f(\underline{x})$, differential entropy can be calculated in closed form only in rare cases. In particular, MDs do not admit a closed-form solution. In that case, one has to rely on the approximations or bounds mentioned above.

C. Fisher Information

Using FI for finding the smoothest continuous density has first been proposed in [8]. It was used as a roughness penalty in maximum likelihood density estimation [9] based on orthogonal Hermite polynomials. Similarly, FI is employed in [28] for wavelet-based density estimation. As a roughness measure, it prevents the density estimate to come too close to the Dirac functions representing the observations. In [12], FI is used for the robust estimation of a location parameter. This method has been extended to minimizing FI over mixtures in [1].

For deriving FI, we start with the so-called score [27, p. 18] given by

$$s(\underline{x};\underline{\theta}) = \frac{\partial}{\partial \underline{\theta}} \log\left(f(\underline{x};\underline{\theta})\right) \tag{7}$$

for some density f depending on a vector parameter $\underline{\theta}$ or equivalently

$$s(\underline{x};\underline{\theta}) = \frac{\partial f(\underline{x};\underline{\theta})}{\partial \underline{\theta}} \frac{1}{f(\underline{x};\underline{\theta})} = \frac{\nabla f(\underline{x};\underline{\theta})}{f(\underline{x};\underline{\theta})} \ . \tag{8}$$

Following [8, p. 29], we define roughness as the difference between the original density $f(\underline{x})$ and a copy $f(\underline{x}+\underline{\theta})$ shifted by a small displacement $\underline{\theta}$. Thus, we have $f(\underline{x};\underline{\theta}) = f(\underline{x}+\underline{\theta})$ and the score is

$$s(\underline{x};\underline{\theta}) = \frac{\partial f(\underline{x}+\underline{\theta})}{\partial \underline{\theta}} \frac{1}{f(\underline{x}+\underline{\theta})} = \frac{\partial f(\underline{x}+\underline{\theta})}{\partial \underline{x}} \frac{1}{f(\underline{x}+\underline{\theta})} \quad . \tag{9}$$

With $\underline{\theta} \to \underline{0}$, the score (8) can be written as [21, p. 2]

$$s(\underline{x}) = \frac{\partial f(\underline{x})}{\partial \underline{x}} \frac{1}{f(\underline{x})} = \frac{\nabla f(\underline{x})}{f(\underline{x})} \tag{10}$$

or the score (7) as

$$s(\underline{x}) = \frac{\partial}{\partial \underline{x}} \log\left(f(\underline{x})\right) \quad . \tag{11}$$

Integrating over the entire domain in the original space of mixtures $\mathcal{M}$ gives the FI [27, p. 15, (1)]

$$I_F^{\mathcal{M}}(f) = \int\limits_{\{\mathbb{R}^D,\, f>0\}} \left| \frac{\nabla f(\underline{x})}{f(\underline{x})} \right|^2 f(\underline{x})\, d\underline{x} \tag{12}$$

or alternatively

$$I_F^{\mathcal{M}}(f) = \int\limits_{\{\mathbb{R}^D,\, f>0\}} \frac{\left|\nabla f(\underline{x})\right|^2}{f(\underline{x})}\, d\underline{x} \quad . \tag{13}$$

We exclude regions where the density approaches zero.

The FI is often simpler to compute than the differential entropy. However, the division by $f(\underline{x})$ in (13) still makes computation difficult especially for mixture densities. In [1], although minimizing FI over mixtures has been considered, computational issues were not addressed.

Summary: FI in the form (12) or (13) cannot be calculated in closed form for MDs because of the division by $f(\underline{x})$.

IV. Closed-form Fisher Information for Mixtures

A. Key Idea

The key idea for obtaining a closed-form expression for the FI of a mixture density is to work in the space of square-root densities, see Fig. 1. In particular, we define RMs that yield the original mixture upon squaring. The desired FI of the original mixture can be calculated in closed form in the RM space. We propose to simultaneously maintain both density representations and perform tandem processing in both spaces. The FI is calculated in the RM space $\mathcal{R}$. The specifications on the pdf $f(\underline{x})$ are calculated in the original mixture space $\mathcal{M}$.

B. Root Densities

Square roots of densities $r(\underline{x}) = \sqrt{f(\underline{x})}$ have already been considered in [9]. The reason for doing so was to ensure that the squared density $f(\underline{x}) = r^2(\underline{x})$ is non-negative. For the same reason, densities based on squared Fourier series are introduced in [2], generalized for multivariate densities in [3], and improved in [24].

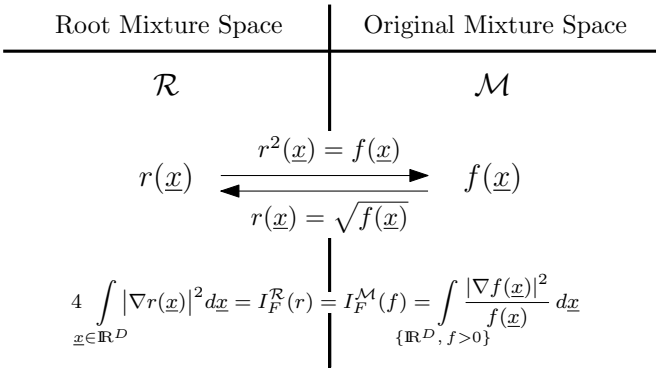

Fig. 1: The two spaces used for calculating least-informative pdfs fulfilling given specifications.

Here, we use square-root pdfs, or in short root densities (RDs), for an entirely different reason. Our considered density representations, e.g., Gaussian mixtures, are non-negative anyway. Instead, we want to obtain a closed-form expression for the FI in (13).

The unit-mass constraint for the pdf $f(\underline{x})$

$$\int\limits_{\mathbb{R}^D} f(\underline{x})\, \mathrm{d}\underline{x} = 1 \tag{14}$$

is equivalent to the constraint

$$\int\limits_{\mathbb{R}^D} r^2(\underline{x})\, \mathrm{d}\underline{x} = 1 \tag{15}$$

for the RD $r(\underline{x})$. This means that $r(\underline{x})$ is restricted to the infinite-dimensional unit sphere S^∞. This manifold is also called the Hilbert sphere [5, p. 2]. When we restrict the RDs $r(\underline{x})$ to be non-negative, i.e., $r(\underline{x}) \geqslant 0\ \forall \underline{x}$, $r(\underline{x})$ is restricted to the "positive orthant" of the Hilbert sphere.

C. Fisher Information for Root Densities

For RDs, the expression for the FI can be simplified. In the univariate case, $D = 1$, with

$$r'(\underline{x}) = \frac{\mathrm{d}}{\mathrm{d}\underline{x}} r(\underline{x}) = \frac{\mathrm{d}}{\mathrm{d}\underline{x}} \sqrt{f(\underline{x})} = -\frac{f'(\underline{x})}{2\sqrt{f(\underline{x})}} \quad , \tag{16}$$

we have

$$\left(r'(\underline{x})\right)^2 = \frac{\left(f'(\underline{x})\right)^2}{4\, f(\underline{x})} \quad . \tag{17}$$

Finally, we can rewrite the FI in the original density space in (13) in the RD space $\mathcal{M}$ as

$$I_F^{\mathcal{R}}(r) = 4 \int\limits_{\underline{x}\in\mathbb{R}^D} \left(r'(\underline{x})\right)^2 d\underline{x} \text{ with } r = \sqrt{f} \quad . \tag{18}$$

In the multivariate case, we have

$$I_F^{\mathcal{R}}(r) = 4 \int\limits_{\underline{x}\in\mathbb{R}^D} \left|\nabla r(\underline{x})\right|^2 d\underline{x} \text{ with } r = \sqrt{f} \quad . \tag{19}$$

These expressions for $I_F^{\mathcal{R}}(r)$ in the RD space $\mathcal{R}$ can be calculated in closed form, e.g., for RMs. In fact, they correspond to the simplified expressions for mean curvature in (5) that are, however, given in the original density space.

D. Root Mixtures

We define RDs in the case of MDs of the form (1) as

$$r(\underline{x}) = \sum_{i=1}^{R} v_i \, r_i(\underline{x}) \ . \tag{20}$$

These square-root MDs will be abbreviated as RMs.

1) Conversion of Root Mixture to Mixture: The conversion of an RM in (20) to a mixture in (1) is unique, always exists, and can be calculated in closed form according to

$$\begin{aligned} f(\underline{x}) = r^2(\underline{x}) &= \left(\sum_{i=1}^{R} v_i \, r_i(\underline{x}) \right) \left(\sum_{i=1}^{R} v_i \, r_i(\underline{x}) \right) \\ &= \sum_{i=1}^{R} \sum_{j=1}^{R} v_i \, v_j \, r_i(\underline{x}) \, r_j(\underline{x}) \ . \end{aligned} \tag{21}$$

This expression contains redundant terms as $r_i(\underline{x}) \, r_j(\underline{x}) = r_j(\underline{x}) \, r_i(\underline{x})$ for $i \neq j$ and can be written as [5]

$$f(\underline{x}) = \sum_{i=1}^{R} \sum_{j=i}^{R} v_i \, v_j \, c_{i,j} \, r_i(\underline{x}) \, r_j(\underline{x}) \tag{22}$$

with

$$c_{i,j} = \begin{cases} 1 \, , \ i = j \\ 2 \, , \ i \neq j \end{cases} \ . \tag{23}$$

The MD $f(\underline{x})$ in (22) contains a total of

$$L = R \cdot (R+1)/2 \tag{24}$$

components [6] and is non-negative by definition (although its weights may be negative). However, without additional constraints on $r(\underline{x})$, the unit mass constraint may be violated.

For deriving specific constraints, we need to consider specific mixture densities. Here, we will consider mixtures of the form (1) with the special case of Gaussian components

$$f_i(\underline{x}) = \frac{1}{\sqrt{|2\pi\Sigma_i|}} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{x}_i)^{\top} \cdot \Sigma_i^{-1} \cdot (\underline{x} - \underline{x}_i) \right\} \tag{25}$$

with $\underline{x} \in \mathbb{R}^D$.

The corresponding Gaussian RM for $\underline{x} \in \mathbb{R}^D$ is given by (20) with Gaussian components

$$r_i(\underline{x}) = \frac{1}{\sqrt{|2\pi P_i|}} \exp\left\{ -\frac{1}{2} (\underline{x} - \underline{\rho}_i)^{\top} \cdot P_i^{-1} \cdot (\underline{x} - \underline{\rho}_i) \right\} \tag{26}$$

with weights v_i, mean vectors $\underline{\rho}_i$, and covariance matrices P_i.

According to (22), the conversion of a Gaussian RM with R components and parameters v_i, $\underline{\rho}_i$, P_i to a GMs results in $L = R \cdot (R+1)/2$ components with w_i, $\underline{x}_i$, Σ_i given by the well-known result of multiyplying two Gaussians, which again yields a Gaussian.

Basic Constraints on Parameters of $r(\underline{x})$: The GM $f(\underline{x})$ resulting from the conversion of the Gaussian RM $r(\underline{x})$ does not necessarily have positive weights that sum to one. For this reason, basic constraints on the parameters of $r(\underline{x})$ in RM space $\mathcal{R}$ are derived from the constraints

$$w_i > 0 \, , \ i = 1{:}L \, , \ \sum_{i=1}^{L} w_i = 1 \tag{27}$$

on the GM $f(\underline{x})$ in the original mixture space $\mathcal{M}$.

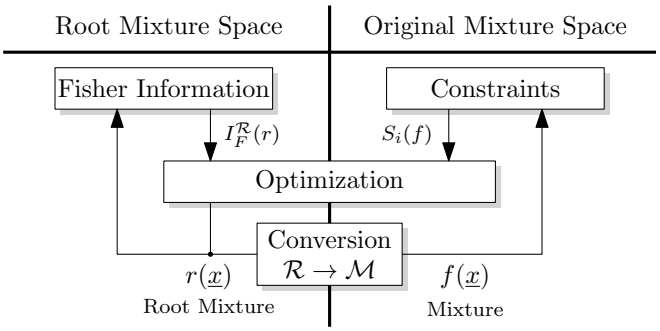

Fig. 2: Overview of optimization via tandem processing in RM space space $\mathcal{R}$ and original mixture space $\mathcal{M}$.

Number of Parameters: The number of parameters of the Gaussian RM $r(\underline{x})$ and the GM $f(\underline{x})$ are equal as the squaring operation in (22) does not add parameters. In the multivariate case, the parameters of $r(\underline{x})$ with R components are given by [7]

$$\underbrace{v_1, v_2, \cdots, v_R}_{R} \, , \ \underbrace{\underline{\rho}_1, \underline{\rho}_2, \cdots, \underline{\rho}_R}_{R \cdot D} \, , \ \underbrace{P_1, P_2, \cdots, P_R}_{R \cdot \frac{D \cdot (D+1)}{2}} \ . \tag{28}$$

With the above mentioned basic constraint we lose one degree of freedom in the weights, so that we end up with

$$(R-1) + R \cdot D + R \cdot \frac{D \cdot (D+1)}{2} = R \cdot \frac{D^2 + 3D + 2}{2} - 1 \tag{29}$$

parameters.

2) Conversion of Mixture to Root Mixture: An RM $r(\underline{x})$ corresponding to a given MD $f(\underline{x})$ does not necessarily exist. This is due to the smaller number of parameters in $r(\underline{x})$ compared to $f(\underline{x})$ so that not every possible $f(\underline{x})$ can be represented by $r(\underline{x})$. In addition, even when it exists, an appropriate RM is non-unique due to, e.g., squared weights. Again, we can exploit small overlaps between distant components when performing the conversion. This reduces the differences in numbers of parameters between the RM and the original mixture.

E. Tandem Processing for Optimization

The RM $r(\underline{x})$ in root mixture space $\mathcal{R}$ is convenient as it allows a closed-form calculation of the FI. In addition, the nonnegativity of the corresponding mixture $f(\underline{x}) = r^2(\underline{x})$ in mixture space $\mathcal{M}$ is automatically guaranteed. However, the unit integral constraint for $f(\underline{x})$ has to be explicitly ensured by an appropriate constraint in mixture space $\mathcal{M}$. Also, the specifications on $f(\underline{x})$ are formulated in mixture space $\mathcal{M}$.

During the optimization procedure, i.e., minimization of FI, see Fig. 2, we maintain both density representations in RM space $\mathcal{R}$ and in mixture space $\mathcal{M}$. This is achieved by directly converting the RM in root mixture space $\mathcal{R}$ to a mixture in the original mixture space $\mathcal{M}$ after any parameter change. FI calculation is performed in RM space $\mathcal{R}$ in tandem with the constraint evaluation in mixture space $\mathcal{M}$.

The optimization problem can be written as

$$\min_{f \in \mathcal{M}} I_F^{\mathcal{M}}(f) \text{ s.t. } S_i(f) = 0 \, , \ S_i \in \mathcal{S} \ , \tag{30}$$

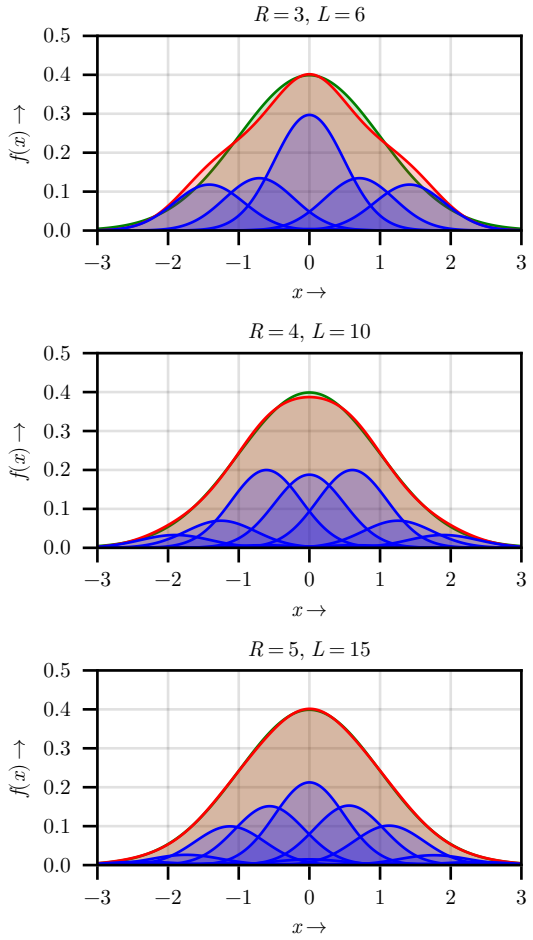

Fig. 3: Red: Smoothest GM $f(\underline{x})$ under specifications zero mean and unit variance in the first example for different numbers of root mixture components R and corresponding mixture components L. Blue: Corresponding mixture components $f_i(\underline{x})$, $i \in 1{:}L$. Green: Gaussian density with zero mean and unit variance as reference.

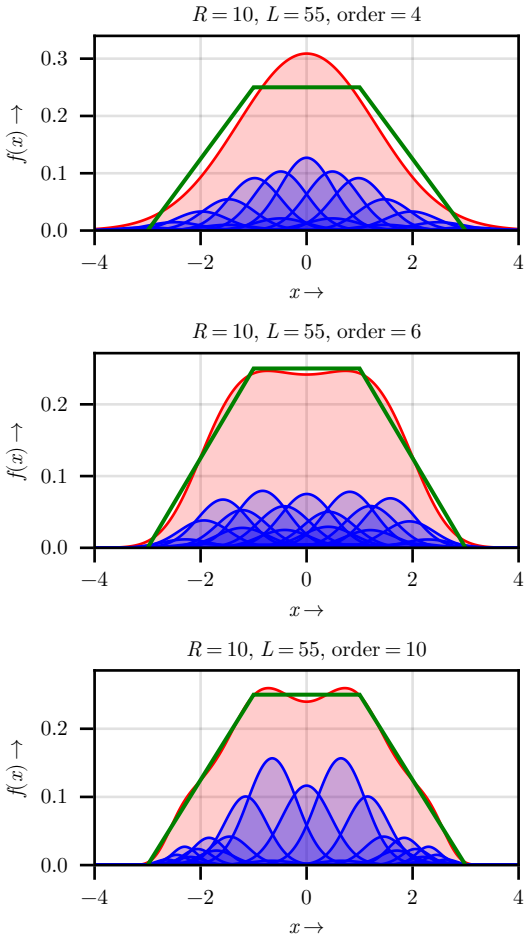

Fig. 4: Red: Smoothest GM $f(\underline{x})$ with moments up to a certain order in the second example for a fixed numbers of root mixture components R and corresponding mixture components L. Green: Underlying (unknown) density function used for calculating the moments. Blue: Corresponding mixture components $f_i(\underline{x})$, $i \in 1{:}L$.

with the FI $I_F^{\mathcal{M}}(f)$ from (12). This is equivalent to

$$\min_{r \in \mathcal{R}} I_F^{\mathcal{R}}(r) \text{ s.t. } f = r^2\,,\ S_i(f) = 0\,,\ S_i \in \mathcal{S}\ , \tag{31}$$

with the FI $I_F^{\mathcal{R}}(r)$ from (18) and $I_F^{\mathcal{M}}(f) = I_F^{\mathcal{R}}(r)$.

For solving the constrained optimization problem, we use an unconstrained optimizer (BFGS) with a penalty function summing up the constraints. The penalty factor is tightened during optimization until the desired constraint accuracy is reached.

V. Examples

We will now give two examples, where we assume a one-dimensional GM $f(x)$ of the form (1) with L components of the form (25).

Example zero mean and unit variance: Besides the constraints on the weights (larger than zero and unit sum), the specifications are only zero mean and unit variance. We know that of all pdfs with given covariance matrix, the Gaussian density minimizes the FI[8] [22] and [23, Lemma 1, p. 184]. Hence, we would expect the GM $f(x)$ to approach a Gaussian shape when the number of components grows. We force the mixture components to have a variance less than 1 to avoid the trivial solution with all components zero mean and variance 1. The results are shown in the original mixture space $\mathcal{M}$ in Fig. 3 for $R \in \{3, 4, 5\}$ RM components. With (22), we obtain $L = R \cdot (R+1)/2 \in \{6, 10, 15\}$ mixture components. For $L = 6$, the number of parameters in the GM $f(x)$ is not sufficient to approach the expected Gaussian density. For $L = 10$, $f(x)$ is closer and for $L = 15$, $f(x)$ is visually indistinguishable from the Gaussian. For $R = 5$ and $L = 15$, it is apparent, that not all components significantly contribute to the density shape as some components are very small.

Example higher-order moments: In this example, we use the weight constraints and moments up to a certain order. For generating the moments, a piecewise linear trapezoidal density function is used, which is shown in green in Fig. 4. The results are shown in Fig. 4 $R = 10$ RM components and $L = 55$ mixture components and for moment orders up 4, 6, and 10. For an increasing order, the calculated mixture density comes closer to the underlying true density.

VI. Conclusions

We derive a closed-form expression for the FI of a GM density $f(\underline{x})$ given in a mixture space $\mathcal{M}$. This is achieved by

introducing an RM space $\mathcal{R}$ and calculating the FI there. The resulting expression consists of a combination of higher-order moments of Gaussian densities, containing nonlinear terms of the component means and covariances. We consider very general specifications of the shape, probability mass distribution, and (central) moments of $f(\underline{x})$. These specifications can also typically be described by nonlinear equations containing component means and covariances. Standard optimization procedures for minimization under equality constraints can be employed to find the least-informative GM density $f(\underline{x})$ under the given specifications.

ENDNOTES

1: By selecting the density $f(\underline{x})$ from a certain class of densities, we inadvertently add information.

2: As a generalization, we can also allow ranges or inequalities for certain specifications. An example would be a mean constraint of the form $E\{\boldsymbol{x}\} \in [-0.2, 0.2]$. Another example is a value constraint of the form $f(x_0) \in [\underline{y}, \overline{y}]$.

3: The underlying pdf $\tilde{f}(\underline{x})$ is mainly a vehicle for explanation. However, in some cases, the goal might indeed be to reconstruct $\tilde{f}(\underline{x})$ from given moments or samples. It is important to note that $\tilde{f}(\underline{x})$ is completely unknown and not used anywhere in the procedure of calculating the desired pdf $f(\underline{x})$.

4: Defining curvature in the multivariate case is much more complicated, see for example [17].

5: Please note that the summations range in $i \in [1, R]$ and $j \in [\boldsymbol{i}, R]$ as we exploit symmetry.

6: When we prespecify the minimum number of components L we desire of a GM in the original mixture space $\mathcal{M}$, the number of RM components in the RM space $\mathcal{R}$ for a full expansion according to (22) are given by

$$R = \left\lceil \frac{\sqrt{8 \cdot L + 1} - 1}{2} \right\rceil \quad , \tag{32}$$

where $\lceil \cdot \rceil$ denotes the next largest integer.

7: The covariance matrices P_i are positive-definite and symmetric. Thus, each P_i is specified by $D \cdot (D+1)/2$ parameters.

8: For a given covariance matrix, the Gaussian density also maximizes the relative entropy [4, p. 411, Example 12.2.1].

REFERENCES

[1] Peter J. Bickel and John R. Collins. "Minimizing Fisher Information over Mixtures of Distributions". *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 45.1 (1983), pp. 1–19.

[2] Dietrich Brunn, Felix Sawo, and Uwe D. Hanebeck. "Efficient Nonlinear Bayesian Estimation based on Fourier Densities". *Proceedings of the 2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2006)*. Heidelberg, Germany, Sept. 2006, pp. 312–322.

[3] Dietrich Brunn, Felix Sawo, and Uwe D. Hanebeck. "Nonlinear Multidimensional Bayesian Estimation with Fourier Densities". *Proceedings of the 2006 IEEE Conference on Decision and Control (CDC 2006)*. San Diego, California, USA, Dec. 2006, pp. 1303–1308.

[4] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. USA: Wiley-Interscience, June 2006.

[5] Xiongtao Dai. *Statistical Inference on the Hilbert Sphere with Application to Random Densities*. Jan. 2, 2021. preprint.

[6] G.A Darbellay and I Vajda. "Entropy Expressions for Multivariate Continuous Distributions". *IEEE Transactions on Information Theory* 46.2 (Mar. 2000), pp. 709–712.

[7] Kenneth Friedman and Abner Shimony. "Jaynes's Maximum Entropy Prescription and Probability Theory". *Journal of Statistical Physics* 3.4 (Dec. 1, 1971), pp. 381–384.

[8] Irving J. Good. "Non-Parametric Roughness Penalty for Probability Densities". *Nature Physical Science* 229.1 (1 Jan. 1971), pp. 29–30.

[9] Irving J. Good and Ray A. Gaskins. "Nonparametric Roughness Penalties for Probability Densities". *Biometrika* 58.2 (Aug. 1, 1971), pp. 255–277.

[10] Uwe D. Hanebeck. "Truncated Moment Problem for Dirac Mixture Densities with Entropy Regularization". *arXiv preprint: Systems and Control (cs.SY)* (Aug. 2014).

[11] Marco F. Huber et al. "On Entropy Approximation for Gaussian Mixture Random Vectors". *Proceedings of the 2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2008)*. Seoul, Republic of Korea, Aug. 2008, pp. 181–188.

[12] Peter J. Huber. "Robust Estimation of a Location Parameter". *The Annals of Mathematical Statistics* 35.1 (1964), pp. 73–101.

[13] Edwin T. Jaynes. "Information Theory and Statistical Mechanics". *Physical Review* 106.4 (May 15, 1957), pp. 620–630.

[14] Edwin T. Jaynes. "Information Theory and Statistical Mechanics. II". *Physical Review* 108.2 (Oct. 15, 1957), pp. 171–190.

[15] Henry J. Landau. "Maximum Entropy and the Moment Problem". *Bulletin (New Series) of the American Mathematical Society* 16.1 (Jan. 1987), pp. 47–77.

[16] Aida V. Lazo and Pushpa N. Rathie. "On the Entropy of Continuous Probability Distributions". *IEEE Transactions on Information Theory* 24.1 (Jan. 1978), pp. 120–122.

[17] John M. Lee. *Riemannian Manifolds: An Introduction to Curvature*. Graduate Texts in Mathematics 176. New York: Springer, 1997. 224 pp.

[18] James Melbourne et al. "The Differential Entropy of Mixtures: New Bounds and Applications". *IEEE Transactions on Information Theory* 68.4 (Apr. 2022), pp. 2123–2146.

[19] Joseph V. Michalowicz, Jonathan M. Nichols, and Frank Bucholtz. "Calculation of Differential Entropy for a Mixed Gaussian Distribution". *Entropy* 10.3 (3 Sept. 2008), pp. 200–206.

[20] Kamyar Moshksar and Amir K. Khandani. "Arbitrarily Tight Bounds on Differential Entropy of Gaussian Mixtures". *IEEE Transactions on Information Theory* 62.6 (June 2016), pp. 3340–3354.

[21] Ágnes Nagy. "Fisher Information and Density Functional Theory". *International Journal of Quantum Chemistry* 122.8 (2022), e26679.

[22] Abbas Pak. "An Alternative Proof For the Minimum Fisher Information of Gaussian Distribution". *Journal of Applied Mathematics, Statistics and Informatics* 14.2 (Dec. 1, 2018), pp. 5–10.

[23] Sangwoo Park, Erchin Serpedin, and Khalid Qaraqe. "Gaussian Assumption: The Least Favorable but the Most Useful". *IEEE Signal Processing Magazine* 30.3 (May 2013), pp. 183–186.

[24] Florian Pfaff, Gerhard Kurz, and Uwe D. Hanebeck. "Multivariate Angular Filtering Using Fourier Series". *Journal of Advances in Information Fusion* 11.2 (Dec. 2016), pp. 206–226.

[25] M. Rangaswamy, D. Weiner, and A. Ozturk. "Non-Gaussian Random Vector Identification Using Spherically Invariant Random Processes". *IEEE Transactions on Aerospace and Electronic Systems* 29.1 (1993), pp. 111–124.

[26] Christian H. Reinsch. "Smoothing by Spline Functions". *Numerische Mathematik* 10.3 (Oct. 1, 1967), pp. 177–183.

[27] Giuseppe Toscani. "Score Functions, Generalized Relative Fisher Information and Applications". *Ricerche di Matematica* 66.1 (June 1, 2017), pp. 15–26.

[28] Marina Vannucci and Brani Vidakovic. "Preventing the Dirac Disaster: Wavelet Based Density Estimation". *Journal of the Italian Statistical Society* 6.2 (Aug. 1, 1997), p. 145.

[29] Chenchen Zhang and Yuan Luo. "Approximating the Differential Entropy of Gaussian Mixtures". *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*. GLOBECOM 2017 - 2017 IEEE Global Communications Conference. Dec. 2017, pp. 1–6.