# Three Approaches to Approximating the Fisher Information Number for Gaussian Mixture Densities

Dominik Prossel and Uwe D. Hanebeck

***Abstract*— The Fisher information number (FIN) has previously been proposed as a regularizer to fit a probability density function to a set of constraints. Especially for mixture densities, this is not straightforward and often a reformulation based on square root densities is used. As it is generally much harder to derive the square root of a mixture than squaring it, this only allows for constraints that can be expressed through the root density's parameters. An important case not covered by this are constraints on individual components of a mixture. This paper proposes three methods to approximate the FIN of mixture models: Gauss-Hermite quadrature, polynomial approximation of the square root function, and direct approximation of the square root density of a pdf. This allows using the FIN for smooth density estimation in situations existing methods cannot handle. The three methods are applied to the problem of kernel density estimation with Gaussian kernels and the results are compared.**

***Index Terms*— Fisher information, density estimation, square root, Gaussian mixture, Gauss-Hermite quadrature**

## I. Introduction

There are many scenarios in information fusion, where only partial information about a probability density is known [1] [2]. This includes density values at some positions, moments of the distribution or the amount of probability mass in certain regions. Reconstructing a continuous probability density function from this information is an ill-posed problem, as more than one feasible solution generally exists. This necessitates using an additional criterion to select one of the solutions. The Fisher information number (FIN) has previously been proposed as a roughness measure to select the smoothest feasible solution [3], [4]. The papers use a Gaussian mixture and a sum of polynomials as density representations and employ a reformulation of the FIN based on the square root of the density to be estimated. This reformulation enables analytic calculation of the FIN for mixture densities, which is an otherwise challenging problem. This approach is suitable for reconstructing smooth densities from "global" information about the density, for example, some of its moments. When working with mixture densities, especially with Gaussian mixtures, some information about the mixture's components might be known like their mean or variance. This case cannot be handled by the methods currently available. The reformulation in terms of the square root density that they use does not allow "local" specifications on components, but only "global" ones on the complete density.

This work has been supported by the Carl Zeiss Foundation under the JuBot project. We also thank the anonymous reviewers for their valuable suggestions.

Dominik Prossel and Uwe D. Hanebeck are with the Intelligent Sensor-Actuator-Systems Laboratory (ISAS), Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany (e-mail: dominik.prossel@kit.edu; uwe.hanebeck@kit.edu)

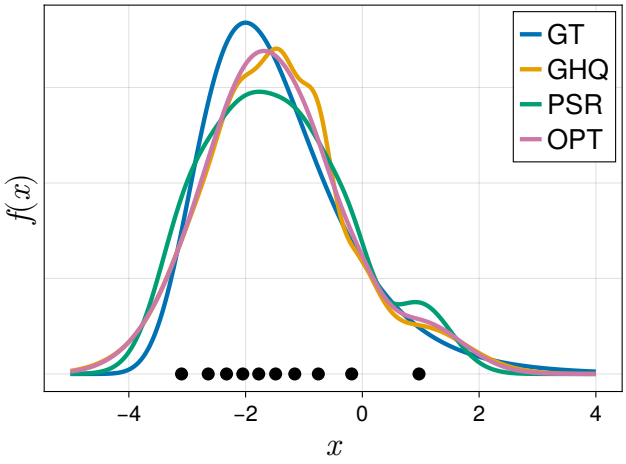



Fig. 1: Densities estimated with the three proposed approximations for the FIN based on eight deterministic samples (black dots) drawn from a Gumbel density (GT). The approximation methods are Gauss-Hermite-Quadrature (GHQ), Polynomial square root (PSR), and Optimization (OPT).

This paper gives a short overview of calculating the FIN of probability distributions and its application to smooth density estimation. Three new approaches to approximate the FIN for mixture densities are proposed. These enable the use of "local" specifications on individual components of a mixture. Their suitability to density estimation is demonstrated through an example application of kernel density estimation, which cannot be solved with the existing methods.

## II. State of the Art

The FIN of a probability density function (pdf) $f(\underline{x})$ in $D$ dimensions with support $\Omega$ is according to [3] defined as

$$I(f(\underline{x})) = \int_{\Omega} \frac{\nabla f(\underline{x})^{\top} \nabla f(\underline{x})}{f(\underline{x})} \, \mathrm{d}\underline{x} \,. \tag{1}$$

This paper assumes that $\Omega$ is the Euclidean space $\mathbb{R}^{\mathbb{D}}$. The FIN has been proposed as a measure of roughness for density estimation based on Hermite polynomials in [4] and was used similarly with Gaussian mixtures in [3] and for wavelet-based density estimation in [5]. In [6] it was used to fit a smooth spline through several given values of the cumulative distribution function.

The integral in (1) can only be solved analytically for some types of distributions, such as Gaussians. The division by $f(\underline{x})$ makes its calculation difficult for mixture densities and polynomials. There is no closed-form solution for Gaussian mixtures available, necessitating numerical integration. The authors of [4] and [3] got around this limitation by rewriting (1) in terms of the square root $r(\underline{x}) = \sqrt{f(\underline{x})}$ of the nonnegative pdf. Making use of

$$\nabla r(\underline{x}) = \nabla\Big(\sqrt{f(\underline{x})}\Big) = \frac{\nabla f(\underline{x})}{2\sqrt{f(\underline{x})}} \ , \tag{2}$$

the FIN can be written as

$$I\big(f(\underline{x})\big) = I\big(r^2(\underline{x})\big) = 4\int_{\mathbb{R}^D} \nabla r(\underline{x})^\top \nabla r(\underline{x})\,\mathrm{d}\underline{x}\,. \tag{3}$$

This eliminates the division by $f(\underline{x})$ in (1), which is the problematic term for mixture densities. The integral in (3) can be solved in closed form for many density representations, most importantly for polynomials and mixture densities, as their derivatives are easy to calculate and square. With this form of the FIN, density estimation can be parameterized in terms of the density's square root $r(\underline{x})$. The actual pdf can then easily be recovered by squaring $r(\underline{x})$. The opposite direction, i.e., calculating the square root $r(\underline{x})$ of a given density $f(\underline{x})$, is more difficult and can generally only be performed approximately.

Because of this limitation, this approach can only be used in cases where all the specifications on $f(\underline{x})$ can be translated into equivalent specifications on $r(\underline{x})$.

An important case not covered by this are constraints on individual mixture components of $f(\underline{x})$ such as fixing their means or covariances or even having certain numbers of components. The components of the mixture density $f(\underline{x})$ are completely defined by squaring $r(\underline{x})$. This confines the possible numbers of components of $f(\underline{x})$ and their location and covariances to certain values so that it is impossible to meet some specifications on $f(\underline{x})$. Another consequence here is, that even though $f(\underline{x})$ has more components than $r(\underline{x})$, both densities have the same number of degrees of freedom. This also makes processing the squared density slightly inefficient, as $f(\underline{x})$ contains redundant information.

## III. Problem Formulation

This paper considers the reconstruction of an unknown multivariate probability density $\tilde{f}(\underline{x})$ from a set of $M$ specifications $\tilde{\mathcal{S}} = \tilde{S}_1(\tilde{f}), \ldots, \tilde{S}_M(\tilde{f})$, which does not completely define $\tilde{f}(\underline{x})$. These specifications can take various forms, such as moments, density values, or the amount of probability mass in certain regions. The estimated density $f(\underline{x})$ should be the least informative pdf fulfilling these specifications. The information content of $f(\underline{x})$ is measured as FIN $I(f(\underline{x}))$ (1). To find the optimal solution $f^*(\underline{x})$, the following general optimization problem is solved

$$\begin{aligned} f^*(\underline{x}) &= \operatorname*{arg\,min}_{f(\underline{x})} I\big(f(\underline{x})\big) \\ \text{s.t.} \quad & \tilde{S}_j(f) = 0 \quad j = 1, \ldots, M\,, \end{aligned} \tag{4}$$

where the specifications $\tilde{S}_j$ are to be satisfied by the estimated density $f(\underline{x})$. To apply (4) to concrete problems, a suitable parametrization of $f(\underline{x})$ has to be selected. It should be able to approximate arbitrary densities and be suitable for further processing like filtering or sampling in multiple dimensions. A common choice fulfilling these requirements is using Gaussian mixture (GM) densities

$$f(\underline{x}; \underline{\theta}_f) = \sum_{i=1}^{L} w_i\, \mathcal{N}(\underline{x}; \underline{\mu}_i, \mathbf{\Sigma}_i) \tag{5}$$

with positive weights $w_i$ summing to one, and Gaussian densities with means $\underline{\mu}_i$ and covariance matrices $\mathbf{\Sigma}_i$. For convenience, these parameters are concatenated into one parameter vector $\underline{\theta}_f$ that specifies the mixture $f(\underline{x}; \underline{\theta}_f)$.

A useful feature of GMs is that the result of addition and multiplication of two GMs is another GM albeit with an increased number of components and not necessarily normalized to integrate to one. The addition operation merges the components of the involved mixtures into one mixture. Multiplication of two GMs can be broken down into a sum of pairwise products between the components of both GMs. There is a well-known formula for the multiplication of two weighted Gaussians $f_1(\underline{x}) = w_1\,\mathcal{N}(\underline{x}; \underline{\mu}_1; \mathbf{\Sigma}_1)$ and $f_2(\underline{x}) = w_2\,\mathcal{N}(\underline{x}; \underline{\mu}_2; \mathbf{\Sigma}_2)$ yielding another Gaussian $f_3(\underline{x}) = f_1(\underline{x}) \cdot f_2(\underline{x}) = w_3\,\mathcal{N}(\underline{x}; \underline{\mu}_3; \mathbf{\Sigma}_3)$ with new parameters

$$\begin{aligned} w_3 &= w_1\, w_2\, \mathcal{N}(\underline{0}; \underline{\mu}_1 - \underline{\mu}_2, \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)\,, \\ \underline{\mu}_3 &= (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}(\mathbf{\Sigma}_2\underline{\mu}_1 + \mathbf{\Sigma}_1\underline{\mu}_2)\,, \\ \mathbf{\Sigma}_3 &= \mathbf{\Sigma}_1(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}\mathbf{\Sigma}_2\,. \end{aligned} \tag{6}$$

Substituting GMs into the optimization problem (4) gives the minimization problem

$$\begin{aligned} \underline{\theta}_f^* &= \operatorname*{arg\,min}_{\underline{\theta}_f} I\big(f(\underline{x}; \underline{\theta}_f)\big) \\ \text{s.t.} \quad & S_k(\underline{\theta}_f) = 0 \quad k = 1, \ldots, 2L+1 \\ & \tilde{S}_j(f) = 0 \quad j = 1, \ldots, M \end{aligned} \tag{7}$$

to find the optimal parameter vector $\underline{\theta}_f^*$. $S_k(\underline{\theta}_f)$ encode the necessary constraints on the parameters of the Gaussian mixture

$$\begin{aligned} &\sum_{i=1}^{L} w_i = 1 \\ &w_i > 0 \quad i = 1, \ldots, L \\ &\Sigma_i \text{ is s.p.d.} \quad i = 1, \ldots, L\ . \end{aligned} \tag{8}$$

As the calculation of the FIN for mixture densities is not straightforward, the paper introduces three approaches to find approximate solutions to (7).

## IV. Approximating the Fisher Information

### A. Gauss-Hermite Quadrature

Quadrature methods in one dimension and cubature methods in multiple dimensions are common ways to perform numerical integration. They work by evaluating the integrand at a specific set of points and calculating a weighted average

of these values. The points and weights are chosen in a way that makes the method exact for certain types of functions. One relatively advanced cubature algorithm is called h-adaptive cubature [7]. It integrates a function over a fixed hypercube domain adaptively refining the discretization to reduce the integration error. This algorithm is well-suited to solve the integral in (1). Its main drawbacks when solving an optimization problem like (7), are its computational cost and differentiability depending on the implementation used.

For these reasons, it is proposed to resort to Gauss-Hermite quadrature instead [8][9]. This method can solve one-dimensional integrals of the form

$$G = \int_{-\infty}^{\infty} g(x)\, e^{-x^2}\, \mathrm{d}x \tag{9}$$

by approximating it as

$$G \approx \sum_{k=1}^{K} \omega_k g(\xi_k)\,. \tag{10}$$

The weights $\omega_k$ and evaluation points $\xi_k$ are typically chosen to be accurate for polynomials of degree $2K-1$ where $K$ points are used for the approximations. Gauss-Hermite quadrature can also be used for integrals involving a Gaussian by transforming the original points to $\tilde{\xi}_k = \mu + \sqrt{2}\sigma\xi_k$ and rescaling the weights $\tilde{\omega}_k = \omega_k/\sqrt{\pi}$. This enables approximating the integral $\tilde{G}$ over a function weighted with a Gaussian as

$$\tilde{G} = \int_{-\infty}^{\infty} g(x)\,\mathcal{N}(x;\mu,\sigma^2)\,\mathrm{d}x \approx \sum_{k=1}^{K} \tilde{\omega}_k\, g(\tilde{\xi}_k)\,. \tag{11}$$

Different methods exist to extend this to multivariate integrals like using the Cartesian product of the samples to yield a grid or transformation to polar coordinates [10]. In this paper, it is chosen to use a scheme similar to the unscented Kalman filter [11] and the filter in [12], where samples are put only on the main axes of the Gaussian. Weights are then appropriately scaled to the increased number of quadrature points. This has the advantage that the number of samples grows linearly with increasing number of dimensions instead of exponentially as with a cartesian grid.

Using (11) and setting

$$g(x) = \frac{\nabla f(\underline{x})^\top \nabla f(\underline{x})}{f(\underline{x})^2}\,, \tag{12}$$

the integral from (1) can then be approximated as

$$I(f(\underline{x})) = \int_{\mathbb{R}^D} \frac{\nabla f(\underline{x})^\top \nabla f(\underline{x})}{f(\underline{x})^2} \sum_{i=1}^{L} w_i\, \mathcal{N}(\underline{x};\underline{\mu}_i,\mathbf{\Sigma}_i)\,\mathrm{d}\underline{x} \tag{13}$$

$$\approx \sum_{i=1}^{L}\sum_{k=1}^{K} w_i\,\tilde{\omega}_k \frac{\nabla f(\underline{\xi}_k)^\top \nabla f(\underline{\xi}_k)}{f(\underline{\xi}_k)^2}\,. \tag{14}$$

### B. Approximating the Square Root Function

A different approach to numerically approximating the integral (1) is to find an approximation to the square root $r(\underline{x})$ that allows using the closed-form solution (3). It is proposed to find a polynomial $p(x)$ of degree $d$ that closely matches the square root function

$$\sqrt{x} \approx p(x) = \sum_{j=0}^{d} c_j\, x^j\,. \tag{15}$$

This polynomial can be found with out-of-the-box function fitting tools. The square root of a probability density $f(\underline{x})$ is then approximately

$$r(\underline{x}) \approx p\big(f(\underline{x})\big)\,. \tag{16}$$

A similar polynomial approximation to $\log(x)$ was proposed in [13] for calculating the entropy of GMs. Polynomials of GMs can be calculated in closed form by repeatedly applying (6), which results in another GM. Depending on the degree of the polynomial and the number of components in $f(\underline{x})$ the number of components in $r(\underline{x})$ can be very high. This makes this method only practical to use with polynomials of relatively low degree. On the other hand, the square root is a notoriously difficult function to approximate with a polynomial, necessitating a tradeoff between accuracy and polynomial degree.

Given the polynomial, the FIN (3) is approximated as

$$\begin{aligned} I\big(f(\underline{x})\big) &= I\Big(p\big(f(\underline{x})\big)^2\Big) \\ &= 4\int_{\mathbb{R}^D} \nabla p\big(f(\underline{x})\big)^\top \nabla p\big(f(\underline{x})\big)\,\mathrm{d}\underline{x} \end{aligned} \tag{17}$$

which can be solved analytically as $p\big(f(\underline{x})\big)$ is a GM.

### C. Simultaneous Optimization

Instead of approximating the square root function $\sqrt{x}$ and applying it to $f(\underline{x})$, the square root $r(\underline{x})$ can also be approximated by directly minimizing the difference between $r^2(\underline{x})$ and $f(\underline{x})$. This results in minimizing the distance measure

$$D\big(f(\underline{x}), r(\underline{x})\big) = \int_{\mathbb{R}^D} \Big(f(\underline{x}) - r(\underline{x})^2\Big)^2\,\mathrm{d}\underline{x}\,. \tag{18}$$

Generally, it is possible to use different types of parametrizations for $r(\underline{x})$ and $f(\underline{x})$. By choosing a Gaussian mixture (GM) representation for both, the distance measure can be calculated in closed form. The expression under the integral in (18) is multiplied out and the resulting sums and products of GMs give again a GM with $T$ components

$$d(\underline{x}) = \Big(f(\underline{x}) - r(\underline{x})^2\Big)^2 = \sum_{t}^{T} v_t \mathcal{N}(\underline{x};\underline{\mu}_t,\mathbf{\Sigma}_t) \tag{19}$$

that can be integrated by summing up the weights $v_t$

$$D\big(f(\underline{x}), r(\underline{x})\big) = \sum_{t=1}^{T} v_t\,. \tag{20}$$

This means that only the weights $v_t$ of $d(\underline{x})$ need to be calculated, which saves some operations compared to computing the complete GM.

The square root density is parametrized as GM $r(\underline{x};\underline{\theta}_r)$ in the following. For a given pdf $f(\underline{x})$, the parameter vector $\underline{\theta}_r^*$

of the best approximation $r(\underline{x};\underline{\theta}_r^*)$ to the true function $r(\underline{x})$ can be found by minimizing

$$\underline{\theta}_r^* = \arg\min_{\underline{\theta}_r} \; D\big(f(\underline{x}), r(\underline{x};\underline{\theta}_r)\big)\,. \tag{21}$$

The FIN of $f(\underline{x})$ can now be approximately calculated in closed-form with (3) by using $r(\underline{x},\underline{\theta}_r^*)$ as a plugin replacement for the true square root density $\sqrt{f(\underline{x})}$.

Incorporating this approximation into the optimization problem (7) is not straightforward. Just substituting it into the objective function would result in a nested optimization problem where the parameters $\underline{\theta}_f$ of $f(\underline{x};\underline{\theta}_f)$ depend on the FIN based on $r(\underline{x};\underline{\theta}_r)$ but the parameters $\underline{\theta}_r$ depend on $\underline{\theta}_f$ to find the square root mixture in the first place. To solve this cyclic dependency the two optimization problems are combined into one with both $\underline{\theta}_f$ and $\underline{\theta}_r$ as optimization variables. This can be done by adding the distance $D\big(f(\underline{x};\underline{\theta}_f), r(\underline{x};\underline{\theta}_r)\big)$ to the objective function yielding

$$\begin{aligned}\underline{\theta}_f^*, \underline{\theta}_r^* = \arg\min_{\underline{\theta}_f,\underline{\theta}_r} & \; I(r^2(\underline{x};\underline{\theta}_f)) + \eta D\big(f(\underline{x};\underline{\theta}_f), r(\underline{x};\underline{\theta}_r)\big) \\ \text{s.t.} \quad & S_k(\underline{\theta}_f) = 0 \quad k = 1,\dots,2L+1 \\ & \tilde{S}_j(f) = 0 \quad j = 1,\dots,M\end{aligned} \tag{22}$$

with a fixed weighting factor $\eta > 0$. $\eta$ has to be chosen large enough to ensure that $f(\underline{x};\underline{\theta}_f)$ and $r^2(\underline{x};\underline{\theta}_r)$ do not drift apart.

Alternatively, (21) can be incorporated into the constraints of (7). A necessary condition at the extreme points of a function is that the gradient is zero at these points

$$\nabla D\big(f(\underline{x};\underline{\theta}_f^*), r(\underline{x};\underline{\theta}_r^*)\big) = \underline{0}\,. \tag{23}$$

This exact condition can be added as a constraint to the optimization problem, ensuring the distance function stays at an optimal value. As this is only a necessary, but not a sufficient condition for a minimum, this can also lead to convergence to a maximum or saddle point of the distance measure. The chances of this happening can be reduced by setting the initial values for $\underline{\theta}_f$ and $\underline{\theta}_r$ to a minimum of the distance or adding a sufficient condition to the constraints.

As the distance measure (18) is symmetric, it is also enough to take the gradient with respect to one of the parameter vectors $\underline{\theta}_f$ and $\underline{\theta}_r$. If one of these gradients is zero, the other one has to be zero as well, as it could not be a minimum otherwise. In practice, it was noticed that the optimizer converges faster when choosing the gradient with respect to $\underline{\theta}_f$. This is also adopted in the final optimization problem

$$\begin{aligned}\underline{\theta}_f^*, \underline{\theta}_r^* = \arg\min_{\underline{\theta}_f,\underline{\theta}_r} & \; I\big(r^2(\underline{x};\underline{\theta}_f)\big) \\ \text{s.t.} \quad & \frac{\partial D\big(f(\underline{x};\underline{\theta}_f), r(\underline{x};\underline{\theta}_r)\big)}{\partial \underline{\theta}_f} = \underline{0} \\ & S_k(\underline{\theta}_f) = 0 \quad k = 1,\dots,2L+1 \\ & \tilde{S}_j(f) = 0 \quad j = 1,\dots,M\end{aligned} \tag{24}$$

While both methods (22) and (24) of rewriting the nested optimization problem work, adding the distance measure into the objective was preferred for the practical experiments, as the resulting problem was much faster to compute.

## V. Example Application

An important application of the methods proposed in this paper is kernel density estimation, where kernels with variable bandwidths are placed at fixed sample positions $\underline{p}_1,\dots,\underline{p}_T$ for $T$ samples. The kernel bandwidths are typically chosen by hand or through heuristics like Silverman's rule of thumb [14]. This can yield undesired results, if the assumptions these methods work with are not fulfilled. Instead, the FIN can be employed to find the kernel bandwidths that yield the overall smoothest density estimate. The optimization problem in (7) can be modified to solve this problem by setting the number of components equal to the number of samples $L = T$, fixing the component weights to $w_i = 1/T$, and the component means to $\underline{\mu}_i = \underline{p}_i$ for $i = 1,\dots,T$. This leaves only the covariance matrices of the components as parameters to be optimized. To simplify computation, especially in high dimensions, the covariance matrices of all components are confined to diagonal matrices.

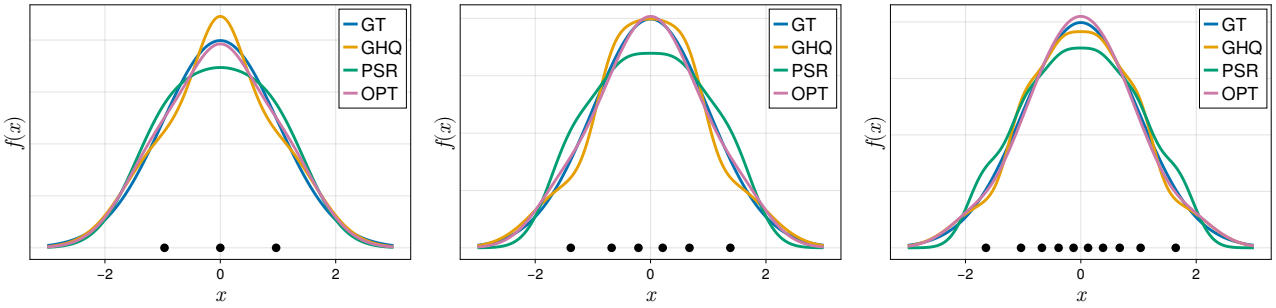
Without further restrictions, minimizing the FIN would result in infinite component covariances. To avoid this undesired effect, the average covariance matrix over all components is limited to a maximum value in each dimension. As the components covariance matrices are diagonal matrices these constraints can easily be expressed as a vector $\underline{C}_{\max}$. This specializes the original optimization problem (7) to

$$\begin{aligned}\underline{\theta}_f^* = \arg\min_{\underline{\theta}_f} & \; I\big(f(\underline{x};\underline{\theta}_f)\big) \\ \text{s.t.} \quad & \underline{\theta}_{f,i} > \underline{0} \quad i = 1,\dots,T \\ & \frac{1}{T}\sum_{i=1}^{T} \underline{\theta}_{f,i} \le \underline{C}_{\max}\;.\end{aligned} \tag{25}$$

with $\underline{\theta}_{f,i}$ containing the diagonal entries of the covariance matrix of the $i$-th component of $f(\underline{x};\underline{\theta}_f)$. The specifications and knowledge about the underlying density are encoded in the fixed component weights and means, and the limited overall covariance.

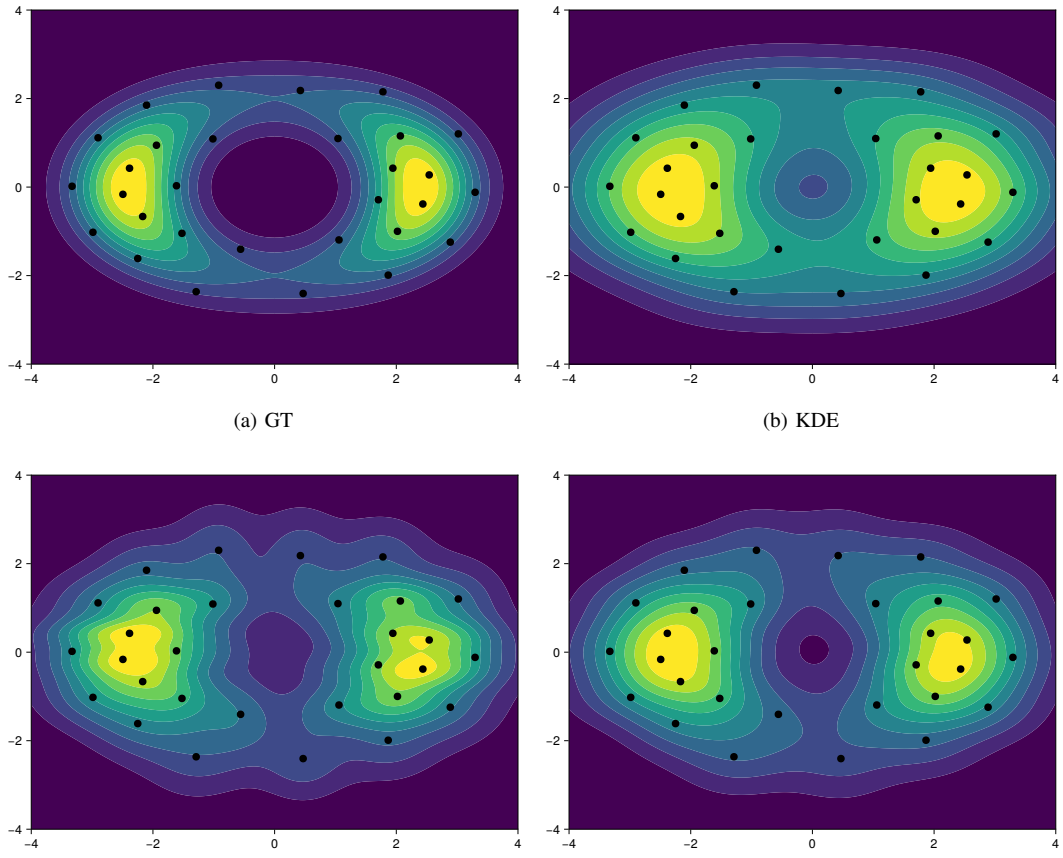
This problem cannot be solved with existing methods employing the FIN as a smoothness measure like [3] or [15]. The reason is that the estimated GM is completely defined through its square root mixture density, which has to have less than $T$ components [3]. It therefore does not have enough degrees of freedom to satisfy all $2T$ constraints on the fixed component weights and means of the GM to be estimated. In this case, the approximations for the FIN presented above can be used.

The three proposed methods to approximate the FIN Gauss-Hermite quadrature (GHQ), polynomial approximation of the square root function (PSR), and the extended optimization problem (OPT) (22) were applied to different one- and two-dimensional examples solving (25). The actual underlying density, that samples were drawn from is shown as ground truth (GT) in the plots, but was not available to the algorithm during estimation. The FIN of the resulting estimates was calculated numerically for comparison using h-adaptive cubature [7].




(a) 3 samples (b) 6 samples (c) 10 samples

Fig. 2: Densities estimated with the three proposed approximations for the FIN based on three/six/ten deterministic samples (black dots) drawn from a Gaussian density.



(a) GT (b) KDE

(c) GHQ (d) OPT

Fig. 3: The underlying density (GT) and densities estimated by kernel density estimation using Silverman's rule of thumb (KDE) and based on two of the proposed approximations for the FIN. The input to the algorithm was 30 deterministic samples (black dots) and constraints on the maximum variance along the coordinate axes.

### A. *One-dimensional examples*

To get an impression of the methods' results, they were applied to a simple one-dimensional problem with three, six, and ten deterministic samples from a standard normal distribution, drawn as described in [16]. The results are shown in Fig. 2. The maximum average component variance was set such that the maximum overall variance was 1.0 for all three experiments. All three methods produce a roughly Gaussian estimate with variance 1.0. The estimate of PSR has a flatter and lower peak than the other two estimates. Most likely this is caused by a bad approximation of the square root function for large values. The FINs of the estimated densities for three samples are 1.12 for GHQ, 1.12 for PSR, and 1.06 for OPT. This is more than the 1.0 expected for a standard normal distribution, which three kernels at the given sample positions cannot exactly represent. This hypothesis is also reinforced by the estimates for six and ten samples that are closer to a normal distribution.

Fig. 1 shows the results for ten deterministic one-dimensional samples from a Gumbel distribution with location parameter $-2.0$ and scale $1.0$. The samples were again generated as in [16]. The estimated densities are reasonably close to the underlying distribution and the peak of PSR is again flatter and lower than with the other methods. The densities have a FIN of $0.76$ for GHQ, $0.86$ for PSR and $0.7$ for OPT. This shows similar results as before, with GHQ and OPT being close, while PSR performing slightly worse.

### B. Two-dimensional example

The methods were also applied to a two-dimensional problem with 30 samples. These were obtained by optimally reducing $10000$ random samples from the density shown in Fig. 3a to 30 samples, as described in [17]. The estimation was only carried out using GHQ and OPT, as PSR proved too computationally expensive for more than 10 samples and did not show promising results in the one-dimensional experiments. A kernel density estimate using Silverman's rule of thumb [14] was also carried out. This rule assumes an underlying Gaussian density to select the kernel bandwidth and is known to give inaccurate results when this is not the case. In Fig. 3b it is evident, that the bandwidth was selected too large, leading to a reconstruction that is too smooth compared to the original density. Both GHQ and OPT estimates better match the shape of the underlying density. Visually, OPT produces a slightly smoother estimate. The difference of the FINs for these methods is minimal, as GHQ yields a FIN of $1.46$ while OPT gives $1.45$.

## VI. Conclusion

This paper introduced three approaches to approximate the FIN for Gaussian mixtures and applied them to a density estimation problem. It shows that some previously infeasible problems can be solved approximately with these methods. Out of the introduced methods, Gauss-Hermite quadrature is identified as giving the best tradeoff between speed and accuracy. The second method of approximating the square root function with a polynomial was found to perform the worst in the investigated scenarios and does not scale well with the number of components. The approximation of the square root density by optimization gives the most accurate results at a high computational cost.

In future research, it would be interesting to identify cases in which having a representation of the square root density could be useful. This could justify the computational cost of the optimization approach, which produces an estimate of the square root density as a currently unused byproduct. While this paper only considers GMs as density representations, the methods proposed can also be applied to other density representations like polynomials or splines. Polynomials, for example, have a very similar structure to GMs with monomials instead of the Gaussians and coefficients in place of weights. This analogy means that they suffer from similar problems as GMs regarding the calculation of the FIN and can very likely profit from the approximations used here.

## References


[1] V. John, I. Angelov, A. A. Öncül, and D. Thévenin, "Techniques for the reconstruction of a distribution from a finite number of its moments," *Chemical Engineering Science*, vol. 62, no. 11, pp. 2890–2904, 2007.

[2] N. Lebaz, A. Cockx, M. Spérandio, and J. Morchain, "Reconstruction of a distribution from a finite number of its moments: A comparative study in the case of depolymerization process," *Computers & Chemical Engineering*, vol. 84, pp. 326–337, 2016.

[3] U. D. Hanebeck, D. Frisch, and D. Prossel, "Closed-Form Information-Theoretic Roughness Measures for Mixture Densities," in *Proceedings of the 2024 American Control Conference (ACC 2024)*, Toronto, Canada, Jul. 2024.

[4] I. J. Good and R. A. Gaskins, "Nonparametric Roughness Penalties for Probability Densities," *Biometrika*, vol. 58, no. 2, pp. 255–277, 1971.

[5] M. Vannucci and B. Vidakovic, "Preventing the Dirac disaster: wavelet based density estimation," *Journal of the Italian Statistical Society*, vol. 6, pp. 145–159, 1997.

[6] P. J. Huber, "Fisher Information and Spline Interpolation," *The Annals of Statistics*, vol. 2, no. 5, pp. 1029–1033, 1974.

[7] A. Genz and A. Malik, "Remarks on Algorithm 006: An Adaptive Algorithm for Numerical Integration over an N-dimensional Rectangular Region," *Journal of Computational and Applied Mathematics*, vol. 6, no. 4, pp. 295–302, 1980.

[8] Q. Liu and D. A. Pierce, "A Note on Gauss—Hermite Quadrature," *Biometrika*, vol. 81, no. 3, pp. 624–629, 1994.

[9] L. N. Trefethen, "Exactness of Quadrature Formulas," *SIAM Review*, vol. 64, no. 1, pp. 132–150, 2022.

[10] P. Jäckel, "A Note on Multivariate Gauss-Hermite Quadrature," *London: ABN-Amro. Re*, 2005. [Online]. Available: `https://api.semanticscholar.org/CorpusID:15184197`.

[11] S. J. Julier and J. K. Uhlmann, "Unscented Filtering and Nonlinear Estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[12] M. F. Huber and U. D. Hanebeck, "Gaussian filter based on deterministic sampling for high quality nonlinear estimation," in *Proceedings of the 17th IFAC World Congress (IFAC 2008)*, vol. 17, Seoul, Republic of Korea, Jul. 2008.

[13] C. Dahlke and J. Pacheco, "On Convergence of Polynomial Approximations to the Gaussian Mixture Entropy," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.

[15] D. Prossel and U. D. Hanebeck, "Spline-Based Density Estimation Minimizing Fisher Information," in *Proceedings of the 27th International Conference on Information Fusion (FUSION 2024)*, Venice, Italy, Jul. 2024.

[16] O. C. Schrempf, D. Brunn, and U. D. Hanebeck, "Density approximation based on dirac mixtures with regard to nonlinear estimation and filtering," in *Proceedings of the 2006 IEEE Conference on Decision and Control (CDC 2006)*, San Diego, California, USA, Dec. 2006.

[17] U. D. Hanebeck, "Optimal reduction of multivariate dirac mixture densities," *at – Automatisierungstechnik*, vol. 63, no. 4, pp. 265–278, Apr. 2015.