Beyond Jensen's Inequality: Speeding Up ML Estimation of Generalized Hyperbolic Distributions

Chenyu Gao
School of Information Science and Technology
ShanghaiTech University
Shanghai, China
gaochy@shanghaitech.edu.cn

Ziping Zhao
School of Information Science and Technology
ShanghaiTech University
Shanghai, China
zipingzhao@shanghaitech.edu.cn

***Abstract*—The generalized hyperbolic (GH) distribution is a highly flexible probability distribution, widely applied across various fields. However, estimating its parameters poses significant challenges. This paper addresses maximum likelihood (ML) estimation for the GH distribution. In the literature, several expectation-maximization (EM)-type algorithms have been proposed, which iteratively solve a surrogate objective, the Q-function—a tractable lower bound to the likelihood function. While these generic EM-type algorithms, grounded in Jensen's inequality, provide a systematic framework, they limit flexibility in algorithm design. In this work, we adopt the minorization-maximization (MM) framework, a more general iterative surrogate maximization approach that subsumes EM as a special case, for ML estimation of the GH distribution. We propose efficient, problem-specific algorithms utilizing novel surrogate functions that offer a provably tighter lower bound to the likelihood than the Q-function, while maintaining closed-form updates. As a result, the proposed algorithms achieve faster convergence with guaranteed convergence. Numerical experiments on synthetic data confirm the superior convergence speed of our algorithms compared to existing methods.**

***Index Terms*—Generalized hyperbolic, parameter estimation, non-convex optimization, expectation-maximization, minorization-maximization.**

I. Introduction

The generalized hyperbolic (GH) distribution is a highly flexible probability distribution, known for its ability to model skewness, kurtosis, and heavy tails [1]. Initially proposed to describe the grain size distributions of wind-blown sand deposits [2], [3], the GH distribution has since been applied to many other fields such as signal processing [4]–[6], machine learning [7]–[10], financial engineering [11], [12], atmospheric engineering [13], and more. As a generic distribution, GH distribution generalizes several well-known distributions, such as Gaussian distribution, Cauchy distribution, skew Cauchy distribution, Student's t distribution, generalized hyperbolic skew Student's t distribution, Laplace distribution, among others, making it a powerful tool for statistical modeling.

The density function of a d-dimensional random variable $\mathbf{x}$ following the multivariate GH distribution is defined as (1),

$$p\left(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\xi}, \lambda, \delta, \gamma\right) = \det\left(2\pi\boldsymbol{\Sigma}\right)^{-\frac{1}{2}} \exp\left(\boldsymbol{\xi}^\top \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right) \cdot \frac{\left(\gamma/\delta\right)^{\lambda}}{K_{\lambda}\left(\delta\gamma\right)} \left(\sqrt{\frac{\delta^2 + \|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}}{\gamma^2 + \|\boldsymbol{\xi}\|^2_{\boldsymbol{\Sigma}^{-1}}}}\right)^{\lambda - \frac{d}{2}} K_{\lambda-\frac{d}{2}}\left(\sqrt{\left(\delta^2 + \|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}\right)\left(\gamma^2 + \|\boldsymbol{\xi}\|^2_{\boldsymbol{\Sigma}^{-1}}\right)}\right) \tag{1}$$

where $\boldsymbol{\mu}$ is the location parameter, $\boldsymbol{\Sigma}$ is the scatter parameter, $\boldsymbol{\xi}$ is the asymmetry parameter, λ and γ are two shape parameters, δ is the scale parameter. K_λ denotes the modified Bessel function of the second kind [14],[1] defined as

$$K_\lambda(x) = \frac{\Gamma\left(\lambda + \frac{1}{2}\right)(2x)^\lambda}{\sqrt{\pi}} \int_0^\infty \frac{\cos t}{\left(t^2 + x^2\right)^{\lambda + 1/2}} \mathrm{d}t.$$

where λ is the order and Γ is the Gamma function. $\|\mathbf{x}\|_{\mathbf{A}}$ is defined as $\sqrt{\mathbf{x}^\top \mathbf{A}\mathbf{x}}$. In this paper, we study the maximum likelihood (ML) estimation method for the GH distribution. Given n independent and identically distributed samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from (1), the ML estimation problem is given by

$$\underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\xi}, \lambda, \delta, \gamma}{\text{maximize}} \quad \sum_{t=1}^{n} \log p(\mathbf{x}_t \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\xi}, \lambda, \delta, \gamma). \tag{2}$$

Solving problem (2) is non-trivial due to the complicated parameter structure in the non-convex objective function.

The expectation-maximization (EM) algorithm is a classical method for ML estimation problem [15], [16]. It is an iterative optimization algorithm consisting of two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, a surrogate function of the likelihood function, known as the Q-function, is derived using Jensen's inequality, a fundamental concept in probability theory. Then, in the M-step, the Q-function is then maximized. Since the GH distribution belongs to the normal mean-variance mixture family, the Q-function function is proven to be jointly concave with respect to all parameters [17]. However, because the Q-function involves multiple parameters, jointly optimizing them presents significant computational challenges. To address this, a variant of EM, called expectation conditional maximization (ECM) [18], simplifies the optimization process by updating one parameter at a time while holding the others fixed—a step referred to as the conditional maximization step (CM-step).

[1] We refer to K_λ as the "Bessel function" for simplicity in the sequel.

An ECM algorithm has been applied to the ML estimation of the GH distribution [19], where parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\xi}$ using closed-form solutions. However, the three scalar parameters λ, δ, and γ lack closed-form solutions and must be optimized using one-dimensional search procedures. When the computation of the E-step is less expensive than the M-step, the multi-cycle ECM (MCECM) becomes a better option. In MCECM, a new Q-function is computed via an E-step before each CM-step, allowing more frequent updates. This approach is expected to result in larger increases in the likelihood function and accelerate convergence. The MCECM method was applied to the ML estimation of the GH distribution in [20]. Despite the improvements offered by ECM and MCECM, one common issue remains: the parameters λ, δ, γ still rely on one-dimensional search procedures. These searches are computationally expensive in practice, as they require repeated evaluation of the Bessel function, which itself is costly. A potential solution is to derive closed-form expressions for these three parameters to accelerate convergence. In [21], the authors developed a modified MCECM algorithm by introducing variational representations for λ, δ, and γ within the Q-function, allowing for closed-form solutions. However, they did not provide any explanation of the algorithm's convergence properties.

In the literature, another algorithm called expectation conditional maximization either (ECME) [22] has been proposed to accelerate the convergence of iterative algorithms. The key idea behind ECME is that directly maximizing the original likelihood function often leads to faster convergence than maximizing a surrogate function. Thus, the approach eliminates the E-step for certain variables, resulting in an algorithm that alternates between maximizing the Q-function and the original likelihood function, depending on the variable being optimized. When applied to the ML estimation of the GH distribution [20], ECME updates $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and $\boldsymbol{\xi}$ through the Q-function with closed-form solutions, while λ, δ, and γ are updated using line-search on the original likelihood function. However, the convergence of ECME is not guaranteed, as the convexity of the likelihood function with respect to λ, δ, and γ remains unknown. In addition to this theoretical concern, as with ECM and the standard MCECM, the performance of ECME in terms of acceleration is case-dependent due to the involvement of line-search, making the algorithm's efficiency vary based on the specific problem at hand.

While the existing generic EM-type algorithms, grounded in Jensen's inequality, provide a systematic framework, they limit flexibility in algorithm design. In this work, we adopt the minorization-maximization (MM) framework, a more general iterative surrogate maximization approach that subsumes EM as a special case, for ML estimation of the GH distribution. In this paper, we employ the block MM (BMM) algorithmic framework to propose a novel algorithm. Under this framework, all the afore-mentioned EM-type algorithms can be viewed as a special case of BMM. In this paper, we make use of the flexibility of the BMM algorithm framework to propose two novel surrogate functions for the variables δ and γ without using the Jensen's inequality. By combining these two surrogate functions, we propose a novel case-by-case surrogation approach, which has not been explored in existing literature. Notably, it can be proved that our proposed surrogate function is a tighter lower bound to the likelihood than the Q-function, while preserving closed-form updates. Comparing with ECME and modified MCECM algorithm [21], which both aim to accelerate the convergence of the algorithm, our algorithm can be seen as a combination of the merits of the two algorithms, in that we would like to seek a tighter surrogate than the Q-function (like what ECME does) and at the same time obtain a closed-form solution (like what the modified MCECM [21] does). We also conduct numerical experiments on synthetic data to confirm the superior convergence speed of our algorithms compared to existing methods.

II. The minorization-maximization algorithm

A general scheme of the block MM framework [23], [24] is first introduced. Given an optimization problem:

$$\underset{\boldsymbol{\theta}\in\Theta}{\text{maximize}} \quad L\left(\boldsymbol{\theta}\right), \tag{4}$$

where the optimization variable $\boldsymbol{\theta}$ is divided into m blocks, as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ with $\boldsymbol{\theta}_i \in \Theta_i$. The product $\Theta = \Theta_1 \times \cdots \times \Theta_m$ is a closed convex set, and $L : \Theta \to \mathbb{R}$ is a continuous function. At each iteration, each block $\boldsymbol{\theta}_i$ is updated in a cyclic order by solving the following problem[2]:

$$\underset{\boldsymbol{\theta}_i\in\boldsymbol{\Theta}_i}{\text{maximize}} \quad S_i\left(\boldsymbol{\theta}_i \middle| \underline{\boldsymbol{\theta}}_1, \ldots, \underline{\boldsymbol{\theta}}_{i-1}, \underline{\boldsymbol{\theta}}_{i+1}, \ldots, \underline{\boldsymbol{\theta}}_m\right) \tag{5}$$

where $S_i\left(\boldsymbol{\theta}_i \middle| \underline{\boldsymbol{\theta}}_{\setminus i}\right)$ with $\underline{\boldsymbol{\theta}}_{\setminus i} \triangleq \left(\underline{\boldsymbol{\theta}}_1, \ldots, \underline{\boldsymbol{\theta}}_{i-1}, \underline{\boldsymbol{\theta}}_{i+1}, \ldots, \underline{\boldsymbol{\theta}}_m\right)$ is a surrogate function, i.e., an lower bound function, of $L(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_i$ satisfying the following conditions:

$$\begin{aligned}
S_i\left(\underline{\boldsymbol{\theta}}_i \middle| \underline{\boldsymbol{\theta}}_{\setminus i}\right) &= L(\underline{\boldsymbol{\theta}}_i | \underline{\boldsymbol{\theta}}_{\setminus i}), \qquad \forall \boldsymbol{\theta}_i \in \Theta_i, \forall \underline{\boldsymbol{\theta}}_{\setminus i} \in \Theta, \forall i. \\
S_i\left(\boldsymbol{\theta}_i \middle| \underline{\boldsymbol{\theta}}_{\setminus i}\right) &\le L(\boldsymbol{\theta}_i | \underline{\boldsymbol{\theta}}_{\setminus i}), \qquad \forall \boldsymbol{\theta}_i \in \Theta_i, \forall \underline{\boldsymbol{\theta}}_{\setminus i} \in \Theta, \forall i.
\end{aligned}$$

In summary, the framework is based on a sequential inexact block coordinate approach, which updates the variable in one block keeping the other blocks fixed. If the surrogate functions S_i are selected properly, the solution to majorized problem (5) could be simpler to obtain than directly solving the original problem (4).

By the assumptions in [23], an optimization method, which is under the BMM framework and has unique solutions for all variables, can converge to the set of stationary point of the original problem. Since the Q-function and the variational representations in the modified MCECM algorithm are the lower bounds of the likelihood function [25], we can establish the convergence for ECM, MCECM and the modified MCECM [21]. In the following, we also follow this framework to construct the novel surrogate functions.

III. Proposed algorithm

Since the modified MCECM algorithm can be interpreted as a BMM, their update rules for four parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\xi}$, λ are also applicable to our proposed BMM. In the following, we focus on construct surrogate functions on the δ and γ sub-problem. The sub-problem is given in (3). The details of the surrogate functions are given in the following.

[2] Throughout this paper, underlined variables denote those whose values are given as constants.

$$\underset{\delta,\,\gamma}{\text{maximize}} \quad n\underline{\lambda}\log\frac{\gamma}{\delta} - \log K_{\underline{\lambda}}\left(\delta\gamma\right) + \frac{1}{2}\left(\underline{\lambda} - \frac{d}{2}\right)\sum_{t=1}^{n}\log\frac{\delta^2 + \|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}}{\gamma^2 + \|\boldsymbol{\xi}\|^2_{\boldsymbol{\Sigma}^{-1}}} + \sum_{t=1}^{n}\log K_{\underline{\lambda}-\frac{d}{2}}\left(\sqrt{\left(\delta^2 + \|\mathbf{x}-\boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}^{-1}}\right)\left(\gamma^2 + \|\boldsymbol{\xi}\|^2_{\boldsymbol{\Sigma}^{-1}}\right)}\right) \tag{3}$$

A. Surrogation 1

Let $\tilde{\underline{\lambda}} = \underline{\lambda} - \frac{d}{2}$, $\tilde{\delta}_t = \sqrt{\delta^2 + \left\|\mathbf{x}_t - \underline{\boldsymbol{\mu}}\right\|^2_{\underline{\boldsymbol{\Sigma}}^{-1}}}$, and $\tilde{\gamma} = \sqrt{\gamma^2 + \left\|\underline{\boldsymbol{\xi}}\right\|^2_{\underline{\boldsymbol{\Sigma}}^{-1}}}$. Then the last two terms in (3) is given by

$$f(\delta,\gamma) = \frac{1}{2}\tilde{\underline{\lambda}}\sum_{t=1}^{n}\log\frac{\tilde{\delta}_t}{\tilde{\gamma}} + \sum_{t=1}^{n}\log K_{\tilde{\underline{\lambda}}}\left(\sqrt{\tilde{\delta}_t\tilde{\gamma}}\right).$$

To find a proper surrogate function for $f(\delta,\gamma)$, we present the following result:

Lemma 1. *Suppose a convex function $g(x)$ can be decomposed as $g(x) = g_1(x) + g_2(x)$, where $g_1(x)$ is a concave function and $g_2(x)$ is convex. Then the surrogate functions $(g_1'(\underline{x}) + g_2'(\underline{x}))x$ is tighter than $g_1(x) + g_2'(\underline{x})x$ for $g(x)$.*

Proof: The proof is given in [25]. ■

Let $f(\delta,\gamma) = f_1(\delta) + f_2(\gamma) + f_3(\delta,\gamma)$, where $f_1(\delta) = \tilde{\underline{\lambda}}\sum_{t=1}^{n}\log\tilde{\delta}_t$, $f_2(\gamma) = -n\tilde{\underline{\lambda}}\log\tilde{\gamma}$, and $f_3(\delta,\gamma) = \sum_{t=1}^{n}\log K_{\tilde{\underline{\lambda}}}\left(\tilde{\delta}_t\tilde{\gamma}\right)$. We observe that $f(\delta,\gamma)$ and $f_3(\delta,\gamma)$ are both concave to γ^2 and δ^2. According to Lemma 1, the convexities of f_1 with respect to γ^2 and f_2 with respect to δ^2 are worthwhile to discuss. Given that the convexities of f_1 and f_2 are influenced by the positivity of $\tilde{\underline{\lambda}}$, we have following two cases:

Case I: When $\tilde{\underline{\lambda}} < 0$, $f_1(\delta)$ convex to δ^2 and $f_2(\gamma)$ is concave to γ^2. Based on Lemma 1, we can derive a tighter lower bound as follows:

$$f(\delta,\gamma) \geq p\delta^2 + f_1(\delta) + u\gamma^2 + \text{const.}, \tag{6}$$

where $p = \frac{\partial f_3(\underline{\delta},\underline{\gamma})}{\partial\delta^2}$ and $u = \frac{\partial f(\underline{\delta},\underline{\gamma})}{\partial\gamma^2}$. Considering the first derivative of the term $f_1(\delta)$ with respect to δ^2, the samples $\mathbf{x}_1,\ldots,\mathbf{x}_N$ are coupled with the variable δ^2. This will result in the loss of a closed-form solution for δ^2. To address this issue, we introduce a result from [24].

Lemma 2. *Function $\sum_{i=1}^{n} a\log f_i(x)$ with $a < 0$ can be minorized as*

$$\sum_{i=1}^{n} a\log f_i(x) \geq \sum_{i=1}^{n} a\log f_i(y) + na\log\left(\frac{\sum_{i=1}^{N} a\frac{f_i(x)}{f_i(y)}}{Na}\right),$$

where the equality is attained when $x = y$. This surrogate function is tighter than directly linearizing the logarithm term.

Based on Lemma 2, we have a lower bound for $f_1(\delta)$ with respect to δ^2

$$f_1(\delta) \geq \frac{n}{2}\tilde{\underline{\lambda}}\log\left(\tau_1\delta^2 + \tau_2\right) + \text{const.}, \tag{7}$$

where $\tau_1 = \sum_{t=1}^{n}\underline{\tilde{\delta}}_t^{-1}$ and $\tau_2 = \sum_{t=1}^{n}\underline{\tilde{\delta}}_t^{-1}\left\|\mathbf{x}_t - \underline{\boldsymbol{\mu}}\right\|^2_{\underline{\boldsymbol{\Sigma}}^{-1}}$.

Case II: When $\tilde{\underline{\lambda}} \geq 0$, $f_1(\delta)$ is concave to δ^2 and $f_2(\gamma)$ is convex to γ^2. Based on Lemma 1, we have the inequality

$$f(\delta,\gamma) \geq v\delta^2 + q\gamma^2 + f_2(\gamma) + \text{const.}, \tag{8}$$

where $v = \frac{\partial f(\underline{\delta},\underline{\gamma})}{\partial\delta^2}$ and $q = \frac{\partial f_3(\underline{\delta},\underline{\gamma})}{\partial\gamma^2}$.

Besides, the Q-function with respect to δ^2 and γ^2 can be interpreted as a linear lower bound on $f(\delta,\gamma)$ as

$$f(\delta,\gamma) \geq u\delta^2 + v\gamma^2 + \text{const.} \tag{9}$$

According to Lemma 1, our proposed surrogate function is tighter than Q-function.

B. Surrogation 2

Due to the term $\log K_{\underline{\lambda}}(\delta\gamma)$, closed-form solutions for δ and γ cannot be obtained. Through a detailed analysis of the Bessel function, we derive a result.

Proposition 3. *Define a function*

$$\phi(\delta,\gamma) = n\underline{\lambda}\log\frac{\gamma}{\delta} - n\log K_{\underline{\lambda}}(\delta\gamma). \tag{10}$$

When $\underline{\lambda} > \frac{1}{2}$, $\phi(\delta,\gamma)$ is convex to δ, and when $\underline{\lambda} \leq -\frac{1}{2}$, $\phi(\delta,\gamma)$ is convex to γ.

Proof: The proof is given in [25]. ■

Hence we have a lower bound for δ on $\underline{\lambda} > \frac{1}{2}$ as

$$\phi(\delta,\gamma) \geq \kappa_1\delta + \text{const}, \tag{11}$$

and a lower bound for γ on $\underline{\lambda} < -\frac{1}{2}$ as

$$\phi(\delta,\gamma) \geq \kappa_2\gamma + \text{const}, \tag{12}$$

where $\kappa_1 = \frac{\partial\phi(\delta,\gamma)}{\partial\delta}$ and $\kappa_2 = \frac{\partial\phi(\delta,\gamma)}{\partial\gamma}$.

Given the lower bound for $\phi(\gamma,\delta)$ in Palmer *et al.*'s paper [21] with respect to γ as follows:

$$\phi(\delta,\gamma) \geq n(\nu + \underline{\lambda})\log\gamma + \text{const.}, \tag{13}$$

where $\nu = \frac{\partial -\log K_{\underline{\lambda}}(\underline{\delta}\underline{\gamma})}{\partial\log(\delta\gamma)}$, due to $\nu + \underline{\lambda} > 0$, this lower bound is concave to γ. In contrast, our lower bound in (12) is linear. Consequently, our lower bound provides a better approximation compared to the formulation in the modified MCECM. Similarly, the lower bound (11) for δ is also tighter than the modified MCECM's formulation.

C. Solving the δ and γ sub-problem

The aforementioned lower bounds work on the different terms in problem (3). Specifically, bounds (6), (7), (8), and (9) are on the last two terms, while (11), (12), and (13) focus on the first two terms. Hence, they can be combined. Since the structure of these lower bounds depends on the value of $\underline{\lambda}$, we identify four distinct cases, each corresponding to specific solutions for δ and γ:

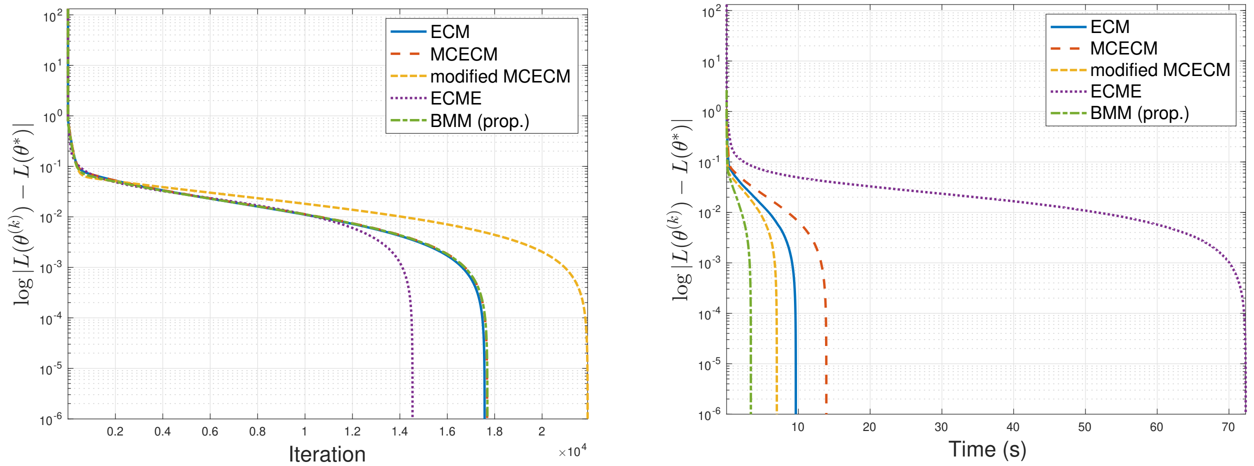

Fig. 1: The convergence of existing and our proposed algorithms. ($d = 10$ and $N = 2000$)

Case I: ($\tilde{\underline{\lambda}} < 0$ and $\underline{\lambda} < -\frac{1}{2}$)

$$\underset{\delta,\,\gamma}{\text{maximize}}\;\; n(\nu - \underline{\lambda}) \log \delta^2 + n\tilde{\underline{\lambda}} \log\left(\tau_1\delta^2 + \tau_2\right) + \kappa_2\gamma \\ + p\delta^2 + v\gamma^2$$

Solution: $\delta^\star$ is solved by the quadratic equation

$$p(\delta^2)^2 + (n(\nu - \underline{\lambda}) + n\tilde{\underline{\lambda}} + p\frac{\tau_2}{\tau_1})\delta^2 + n(\nu - \underline{\lambda})\frac{\tau_2}{\tau_1} = 0, \quad (14)$$

and $\gamma^\star = -\frac{\kappa_2}{v}$.

Case II: ($\tilde{\underline{\lambda}} < 0$ and $-\frac{1}{2} \le \underline{\lambda} \le \frac{1}{2}$)

$$\underset{\delta,\,\gamma}{\text{maximize}}\;\; n(\nu \log(\delta^2\gamma^2) + \underline{\lambda} \log \frac{\gamma^2}{\delta^2}) + N\tilde{\underline{\lambda}} \log\left(\tau_1\delta^2 + \tau_2\right) \\ + p\delta^2 + v\gamma^2$$

Solution: $\delta^\star$ is solved by the quadratic equation (14) and $\gamma^\star = \sqrt{-\frac{\nu+\underline{\lambda}}{v}}$

Case III: ($\tilde{\underline{\lambda}} < 0$ and $\underline{\lambda} > \frac{1}{2}$)

$$\underset{\delta,\,\gamma}{\text{maximize}}\;\; n(\kappa_2\delta + (\nu + \underline{\lambda}) \log \gamma^2) + u\delta^2 + v\gamma^2$$

Solution: $\delta^\star = -\frac{\kappa_2}{v}$ and $\gamma^\star = \sqrt{-\frac{\nu+\underline{\lambda}}{u}}$.

Case IV: ($\tilde{\underline{\lambda}} \ge 0$ and $\underline{\lambda} > \frac{1}{2}$)

$$\underset{\delta,\,\gamma}{\text{maximize}}\;\; n(\kappa_2\delta(\nu + \underline{\lambda}) + \log \gamma^2 - \tilde{\underline{\lambda}} \log \tilde{\gamma}) + u\delta^2 + q\gamma^2$$

Solution: $\delta^\star = -\frac{\kappa_2}{v}$ and $\gamma^\star$ is solved by a quadratic function

$$q(\gamma^2)^2 + (n(\nu+\underline{\lambda}) - \tilde{\underline{\lambda}} + q\left\|\underline{\boldsymbol{\xi}}\right\|^2_{\underline{\boldsymbol{\Sigma}}^{-1}})\delta^2 + n(\nu+\underline{\lambda})\left\|\underline{\boldsymbol{\xi}}\right\|^2_{\underline{\boldsymbol{\Sigma}}^{-1}} = 0. \quad (15)$$

Note that each of quadratic equations (14) and (15) has only one positive solution. These solutions correspond to the global maximizers of their respective optimization problems.

IV. Experiments

In this section, we carry out numerical simulations using both synthetic and real-world data to evaluate the performance of our proposed algorithm with comparison to the existing methods. The numerical simulations are performed on a personal computer with a 3.3 GHz Intel Xeon W CPU.

The algorithms under comparisons include ECM, MCECM, modified MCECM, ECME algorithm and our proposed algorithm. To ensure a fair comparison of time efficiency across different methods, all algorithms are implemented using MATLAB.

We first generate a set of synthetic data with 2000 samples, 10 dimensions, and true scalar parameters $\lambda = -3$, $\gamma = 3$, and $\delta = 2$. Since our proposed algorithm focus on the estimation of scalar parameters, we only discuss the selections of true parameters.

Figure 1 illustrates the convergence rates of our proposed algorithm compared to existing methods. The y-axis represents the difference between the objective value at each iteration and the final converged value, and the x-axis is iteration and CPU time, respectively. In this figure, our algorithm demonstrates the lowest CPU time among the compared methods.

V. Conclusions

In this paper, we tackle the challenges of estimating parameters for the multivariate generalized hyperbolic (GH) distribution. We interpret the existing methods into the EM-type and give the modified MCECM method a convergence guarantee. Then we propose a novel EM-type algorithm which further explore the structure of ML estimation problem and analysis the properties of the Bessel function. The experiments demonstrate that our algorithm costs less time for convergence. Since GH encompasses many distributions, our proposed algorithm can also extend to estimate these distributions.

It is noteworthy that our algorithms have a similar concept with the alternating ECM algorithm [26], [27]. This approach constructs Q-functions that differ from those in traditional EM algorithms by using flexible data augmentation schemes [28] for various blocks. Although we have not explicitly derived the hidden variables corresponding to our proposed lower bounds, we hypothesize that such data augmentation schemes exist.

References

[1] K. Prause *et al.*, "The generalized hyperbolic model: Estimation, financial derivatives, and risk measures," Ph.D. dissertation, Universitüt Freiburg, 1999.

[2] O. Barndorff-Nielsen, "Exponentially decreasing distributions for the logarithm of particle size," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 353, no. 1674, pp. 401–419, 1977.

[3] O. E. Barndorff-Nielsen, P. Blæsild, J. L. Jensen, and M. Sørensen, *The Fascination of Sand*. Springer, 1985.

[4] J. Nowicka-Zagrajek and R. Weron, "Modeling electricity loads in California: ARMA models with hyperbolic noise," *Signal Processing*, vol. 82, no. 12, pp. 1903–1915, 2002.

[5] S. Cumani, "On the distribution of speaker verification scores: Generative models for unsupervised calibration," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 547–562, 2021.

[6] M. C. Amrouche, H. Carfantan, and J. Idier, "A partially collapsed Gibbs sampler for unsupervised nonnegative sparse signal restoration," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5519–5523.

[7] H. Snoussi and J. Idier, "Bayesian blind separation of generalized hyperbolic processes in noisy and underdeterminate mixtures," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3257–3269, 2006.

[8] K. Morris and P. D. McNicholas, "Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures," *Computational Statistics & Data Analysis*, vol. 97, pp. 133–150, 2016.

[9] C. Tortora, P. D. Mcnicholas, and R. P. Browne, "A mixture of generalized hyperbolic factor analyzers," *Advances in Data Analysis and Classification*, vol. 10, no. 4, pp. 423–440, 2016.

[10] L. Cheng, Z. Chen, Q. Shi, Y.-C. Wu, and S. Theodoridis, "Towards flexible sparsity-aware modeling: Automatic tensor rank learning using the generalized hyperbolic prior," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1834–1849, 2022.

[11] E. Eberlein and U. Keller, "Hyperbolic distributions in finance," *Bernoulli*, vol. 1, no. 3, pp. 281–299, 1995.

[12] M. Predota, "On European and Asian option pricing in the generalized hyperbolic model," *European Journal of Applied Mathematics*, vol. 16, no. 1, pp. 111–144, 2005.

[13] K. Cugerone and C. De Michele, "Investigating raindrop size distributions in the (L-)skewness–(L-)kurtosis plane," *Quarterly Journal of the Royal Meteorological Society*, vol. 143, no. 704, pp. 1303–1312, 2017.

[14] G. N. Watson, *A Treatise on the Theory of Bessel Functions*. The University Press, 1922, vol. 2.

[15] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[16] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. John Wiley & Sons, 2007.

[17] O. Barndorff-Nielsen, J. Kent, M. Sørensen, and M. Sorensen, "Normal variance-mean mixtures and z distributions," *International Statistical Review*, vol. 50, no. 2, p. 145, 1982.

[18] X.-L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ecm algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[19] R. S. Protassov, "EM-based maximum likelihood parameter estimation for multivariate generalized hyperbolic distributions with fixed λ," *Statistics and Computing*, vol. 14, pp. 67–77, 2004.

[20] A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2005.

[21] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "An EM algorithm for maximum likelihood estimation of Barndorff-Nielsen's generalized hyperbolic distribution," in *2016 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2016, pp. 1–4.

[22] C. Liu and D. B. Rubin, "The ECME algorithm: A simple extension of EM and ECM with faster monotone convergence," *Biometrika*, vol. 81, no. 4, pp. 633–648, 1994.

[23] M. Razaviyayn, M. Hong, and Z. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, no. 2, pp. 1126–1153, 2013.

[24] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2016.

[25] [Online]. Available: https://www.ncvxopt.com/pubs/GaoZhaoGH

[26] J. Fessler and A. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2664–2677, Oct./1994.

[27] X.-L. Meng and D. Van Dyk, "The EM algorithm—an old folk-song sung to a fast new tune," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 59, no. 3, pp. 511–567, 1997.

[28] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, Mar. 2001.

Supplementary Materials for "Beyond Jensen's Inequality: Speeding Up ML Estimation of Generalized Hyperbolic Distributions"

Chenyu Gao and Ziping Zhao

Appendix A
Proof the variational representations in the modified MCECM as lower bounds

Proof: Let $f_1(x)$ and $f_2(x)$ be differentiable functions. A sufficient condition for $f_2(x)$ is a lower bound of $f_1(x)$ is that

$$\begin{cases} g(x) > 0 & x > \underline{x}, \\ g(x) < 0 & x < \underline{x}. \end{cases} \tag{1}$$

where $g(x) = f_1'(x) - f_2'(x)$.

We begin by examining the lower bound for λ. Let $f_1(\lambda) = -\log K_\lambda(\underline{\delta\gamma})$ and $f_2(\lambda) = \frac{\partial - \log K_{\underline{\lambda}}(\underline{\delta\gamma})}{\partial \underline{\lambda}^2}\lambda^2$. The difference between their first derivatives with respect to λ is

$$\begin{aligned} g_1(\lambda) &= f_1'(\lambda) - f_2'(\lambda) \\ &= \frac{\partial - \log K_\lambda(\underline{\delta\gamma})}{\partial \lambda} - \frac{\partial - \log K_{\underline{\lambda}}(\underline{\delta\gamma})}{\partial \underline{\lambda}^2} 2\lambda \\ &= 2\lambda \left(\frac{\partial - \log K_\lambda(\underline{\delta\gamma})}{\partial \lambda} \frac{1}{2\lambda} - \frac{\partial - \log K_{\underline{\lambda}}(\underline{\delta\gamma})}{\partial \underline{\lambda}^2} \right) \\ &= 2\lambda \left(\frac{\partial - \log K_\lambda(\underline{\delta\gamma})}{\partial \lambda^2} - \frac{\partial - \log K_{\underline{\lambda}}(\underline{\delta\gamma})}{\partial \underline{\lambda}^2} \right), \end{aligned}$$

with $g_1(\underline{\lambda}) = 0$.

Next, we analyze the sign of $g_1(\lambda)$ for $\lambda > 0$ and $\underline{\lambda} > 0$. Referring to [1, Theorem 1], the function $-\log K_\lambda(x)$ is convex with respect to λ^2 on $\lambda > 0$. This convexity implies that $\frac{\partial \log K_\lambda(\underline{\delta\gamma})}{\partial \lambda^2}$ is increasing. Consequently, $g_1(\lambda)$ is monotonically increasing, so that the condition (1) is met. Therefore, $f_2(\lambda)$ is the lower bound of $f_1(\lambda)$ for $\lambda > 0$ and $\underline{\lambda} > 0$. Since both $f_1(\lambda)$ and $f_2(\lambda)$ are even, the lower bound extends to all $\lambda \neq 0$ and $\underline{\lambda} \neq 0$.

Next, we consider the lower bound for $\delta\gamma$. Let $f_1(\delta\gamma) = -\log K_{\underline{\lambda}}\left(\delta\gamma\right)$ and $f_2(\delta\gamma) = \frac{1}{2}\frac{\partial - \log K_{\underline{\lambda}}\left(\underline{\delta\gamma}\right)}{\partial \log(\underline{\delta\gamma})} \log(\delta^2\gamma^2)$. Then the the first derivative of $f_1(\delta\gamma) - f_2(\delta\gamma)$ with respect to $\delta\gamma$ is given by[1]

$$\begin{aligned} g_2(\delta\gamma) &= f_1'(\delta\gamma) - f_2'(\delta\gamma) \\ &= \frac{\partial - \log K_{\underline{\lambda}}\left(\delta\gamma\right)}{\partial \delta\gamma} - \frac{\partial - \log K_{\underline{\lambda}}\left(\underline{\delta\gamma}\right)}{\partial \log(\underline{\delta\gamma})} \frac{1}{\delta\gamma} \\ &= \frac{1}{\delta\gamma} \left(\delta\gamma \frac{\partial - \log K_{\underline{\lambda}}\left(\delta\gamma\right)}{\partial \delta\gamma} - \underline{\delta\gamma} \frac{\partial - \log K_{\underline{\lambda}}\left(\underline{\delta\gamma}\right)}{\partial \underline{\delta\gamma}} \right) \\ &= -\frac{1}{\delta\gamma} \left(\delta\gamma \frac{K_{\underline{\lambda}}'\left(\delta\gamma\right)}{K_{\underline{\lambda}}\left(\delta\gamma\right)} - \underline{\delta\gamma} \frac{K_{\underline{\lambda}}'\left(\underline{\delta\gamma}\right)}{K_{\underline{\lambda}}\left(\underline{\delta\gamma}\right)} \right). \end{aligned}$$

Since the function $x\frac{K_{\underline{\lambda}}'(x)}{K_{\underline{\lambda}}(x)}$ is strictly decreasing on $x > 0$, the condition (1) is satisfied and so that $f_2(\delta\gamma)$ is a lower bound of $f_1(\delta\gamma)$. ■

Appendix B
Proof of Lemma 1

Proof: Given that $g(x)$ and $g_2(x)$ are convex functions, and $g_1(x)$ is concave, we can establish two surrogate functions as follows:

$$g(x) \geq g(\underline{x}) + \big(g_1'(\underline{x}) + g_2'(\underline{x})\big)(x - \underline{x}),$$

and

$$g(x) \geq g_1(x) + g_2(\underline{x}) + g_2'(\underline{x})(x - \underline{x}).$$

[1] Throughout this paper, $\frac{\partial}{\partial x}K_\lambda(x)$ is denoted as $K_\lambda'(x)$.

Let $r_1(x)$ and $r_2(x)$ denote the right-hand sides of the first and second inequalities, respectively. To determine which is the tighter lower bound, we examine the difference between them:

$$r_1(x) - r_2(x) = g_1(\underline{x}) - g_1(x) + g_1'(\underline{x})(x - \underline{x}).$$

Since $g_1(x)$ is concave, it satisfies the inequality:

$$g_1(x) \leq g_1(\underline{x}) + g_1(\underline{x})(x - \underline{x}).$$

Hence, we have $r_1(x) - r_2(x) \geq 0$ so that $r_1(x)$ is the tighter surrogation compared to $r_2(x)$. ■

APPENDIX C
PROOF OF PROPOSITION 3

Proof: Without loss of generality, we assume $n = 1$ for this proof. Consider the function

$$\phi(\delta, \gamma) = \underline{\lambda} \log \frac{\gamma}{\delta} - \log K_{\underline{\lambda}}(\delta\gamma).$$

We define that $\phi(\delta) = \phi(\delta, \gamma = \underline{\gamma})$ and $\phi(\gamma) = \phi(\delta = \underline{\delta}, \gamma)$. For $\underline{\lambda} > \frac{1}{2}$, the second derivative for $\phi(\delta)$ can be computed as

$$\phi''(\delta) = \frac{\underline{\lambda}}{\delta^2} - \frac{K''_{\underline{\lambda}}\Big(\delta\underline{\gamma}\Big)}{K_{\underline{\lambda}}\Big(\delta\underline{\gamma}\Big)}\underline{\gamma}^2 + \left(\frac{K'_{\underline{\lambda}}\Big(\delta\underline{\gamma}\Big)}{K_{\underline{\lambda}}\Big(\delta\underline{\gamma}\Big)}\underline{\gamma}\right)^2.$$

Based on the expression of the second derivative of Bessel function

$$K''_\lambda(x) = (1 + \frac{\lambda^2}{x^2})K_\lambda(x) - \frac{K'_\lambda(x)}{x}, \tag{2}$$

we have

$$\phi''(\delta) = -\frac{\underline{\lambda} + \delta^2\underline{\gamma}^2 + \underline{\lambda}^2}{\delta^2} + \frac{K'_{\underline{\lambda}}\Big(\delta\underline{\gamma}\Big)}{K_{\underline{\lambda}}\Big(\delta\underline{\gamma}\Big)}\frac{\delta}{\underline{\gamma}} + \left(\frac{K'_{\underline{\lambda}}\Big(\delta\underline{\gamma}\Big)}{K_{\underline{\lambda}}\Big(\delta\underline{\gamma}\Big)}\underline{\gamma}\right)^2.$$

For the function $\frac{K'_\lambda(x)}{K_\lambda(x)}$, based on the bounds for $|\lambda| > \frac{1}{2}$ as established in [2], we have

$$-\frac{1 + \sqrt{(|\lambda| - 1)^2 + x^2}}{x} \leq \frac{K'_\lambda(x)}{K_\lambda(x)} \leq -\frac{\frac{1}{2} + \sqrt{(|\lambda| - \frac{1}{2})^2 + x^2}}{x}. \tag{3}$$

Applying these bounds to $\phi''(\delta)$, we obtain

$$\begin{aligned}\phi''(\delta) &> -\frac{-\underline{\lambda} + \delta^2\underline{\gamma}^2 + \underline{\lambda}^2}{\delta^2} - \frac{\underline{\gamma}}{\delta}\frac{1 + \sqrt{(|\underline{\lambda}| - 1)^2 + \delta^2\underline{\gamma}^2}}{\delta\underline{\gamma}} \\ &\quad + \left(-\frac{\frac{1}{2} + \sqrt{(|\underline{\lambda}| - \frac{1}{2})^2 + \delta^2\underline{\gamma}^2}}{\delta\underline{\gamma}}\underline{\gamma}\right)^2 \\ &= -\frac{-2\underline{\lambda} + \frac{1}{2} + h(\delta, \underline{\gamma})}{\delta^2},\end{aligned}$$

where

$$\begin{aligned}h(\delta, \gamma) &= \sqrt{(|\underline{\lambda}| - 1)^2 + \delta^2\gamma^2} - \sqrt{(|\underline{\lambda}| - \frac{1}{2})^2 + \delta^2\gamma^2} \\ &= \frac{-|\underline{\lambda}| + \frac{3}{4}}{\sqrt{(|\underline{\lambda}| - 1)^2 + \delta^2\gamma^2} + \sqrt{(|\underline{\lambda}| - \frac{1}{2})^2 + \delta^2\gamma^2}}.\end{aligned} \tag{4}$$

To establish the convexity of $\phi(\delta)$, we consider two cases:

1) When $\underline{\lambda} > \frac{3}{4}$, we have $h(\delta, \underline{\gamma}) < 0$ so that $\phi''(\delta) > 0$.
2) When $\frac{1}{2} < \underline{\lambda} < \frac{3}{4}$, we have $h(\delta, \gamma)$ is strictly decreasing on both δ and γ, i.e., $0 < h(\delta, \underline{\gamma}) < h(0, \underline{\gamma}) = -2|\underline{\lambda}| + \frac{3}{2}$. Hence, we have
$$\phi''(\delta) > -\frac{-2\underline{\lambda} + \frac{1}{2} + h(\delta, \underline{\gamma})}{\delta^2} > -\frac{-4\underline{\lambda} + 2}{\delta^2} > 0.$$

Therefore, $\phi(\delta)$ is a convex function.

Next, consider $\underline{\lambda} < -\frac{1}{2}$. The second derivative for $\phi(\gamma)$ is computed as

$$\begin{aligned}\phi''(\gamma) &= -\frac{\underline{\lambda}}{\gamma^2} - \frac{K''_{\underline{\lambda}}\Big(\underline{\delta}\gamma\Big)}{K_{\underline{\lambda}}\Big(\underline{\delta}\gamma\Big)}\underline{\delta}^2 + \left(\frac{K'_{\underline{\lambda}}\Big(\underline{\delta}\gamma\Big)}{K_{\underline{\lambda}}\Big(\underline{\delta}\gamma\Big)}\underline{\delta}\right)^2 \\ &> -\frac{2\underline{\lambda} + \frac{1}{2} + h(\underline{\delta}, \gamma)}{\gamma^2}.\end{aligned}$$

Therefore, we can similarly obtain that $\phi''(\gamma)$ is convex for $\gamma > 0$ and $\underline{\lambda} > \frac{1}{2}$. ■

References

[1] Á. Baricz and S. Ponnusamy, "On Turán type inequalities for modified Bessel functions," *Proceedings of the American Mathematical Society*, vol. 141, no. 2, pp. 523–532, 2012.

[2] J. Segura, "Simple bounds with best possible accuracy for ratios of modified Bessel functions," *Journal of Mathematical Analysis and Applications*, vol. 526, no. 1, p. 127211, 2023.