GLOBALLY CONVERGENT ALGORITHMS FOR LEARNING MULTIVARIATE GENERALIZED GAUSSIAN DISTRIBUTIONS

Bin Wang[♯], Huanyu Zhang[♯], Ziping Zhao[♯], and Ying Sun[†]

[♯]School of Info. Sci. and Tech., ShanghaiTech University, Shanghai 201210, China
[†]School of Elec. Eng. and Comp. Sci., The Pennsylvania State University, PA 16802, USA
Email: {wangbin, zhanghy4, zhaoziping}@shanghaitech.edu.cn, ysun@psu.edu

2021 IEEE Statistical Signal Processing Workshop (SSP) | 978-1-7281-5767-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SSP49050.2021.9513857

ABSTRACT

The multivariate generalized Gaussian distribution has been used intensively in various data analytics fields. Due to its flexibility in modeling different distributions, developing efficient methods to learn the model parameters has attracted lots of attentions. Existing algorithms including the popular fixed-point algorithms focus on learning the shape parameters and scatter matrices, but convergence is only established when the shape parameters are taken as given. When coupled with the shape parameters, convergence properties of the existing alternating algorithms remain unknown. In this paper, globally convergent algorithms based on the block majorization minimization method are proposed to jointly learn all the model parameters in the maximum likelihood estimation setting. The negative log-likelihood function w.r.t. the shape parameter is proved to be strictly convex, which to our best knowledge is the first result of this kind in the literature. Superior performance of the proposed algorithms are validated numerically based on synthetic data with comparisons to existing methods.

I. INTRODUCTION

The multivariate generalized Gaussian distribution (MGGD) [1], a.k.a. multivariate exponential power distribution [2], has aroused a great interest in the signal processing community. Since it is flexible and powerful in modeling data with different distribution properties, they have been intensively used in many signal processing applications including image denoising [3], image/video segmentation [4], video coding [5], computer vision [6], ultrawide bandwidth communications [7], biomedical signal processing [8], radar signal processing [9], [10], and financial signal processing [11], [12].

Generally speaking, MGGD belongs to the family of elliptical distributions. The probability density function of an MGGD [13] for $\mathbf{x} \in \mathbb{R}^p$ is given as follows:[1]

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma},\beta) = \frac{\beta\Gamma\left(\frac{p}{2}\right)}{2^{\frac{p}{2\beta}}(\pi)^{\frac{p}{2}}\Gamma\left(\frac{p}{2\beta}\right)} \frac{1}{\det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \times \exp\left[-\frac{\left((\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{\beta}}{2}\right], \tag{1}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the location vector, $\boldsymbol{\Sigma} \in \mathbb{S}_{++}^{p\times p}$ is the scatter matrix (a.k.a. scale/dispersion matrix), $\beta \in (0, +\infty)$ is the shape parameter, and $\Gamma(\cdot)$ denotes the gamma function defined by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$. The shape parameter β characterizes the peakedness and spread of the MGGDs [14]. When $\beta = 1$, it corresponds to the multivariate Gaussian distribution. The marginal distribution of the MGGD is more peaky with heavy tails if $\beta < 1$ (Specifically, it becomes a multivariate Laplacian distribution when $\beta = \frac{1}{2}$.) and less peaky with light tails if $\beta > 1$ (When $\beta \to \infty$, it tends to converge to a multivariate uniform distribution.). Motivated by practical applications like in image processing, array signal processing, and financial signal processing, the cases when $\beta \in (0, 1]$ are usually more interesting [15].

Considering the broad applications of MGGDs, learning the model parameters efficiently and accurately becomes an important task. In the literature, several methods have been proposed to solve the MGGD parameter learning problem based on methods like the maximum likelihood estimation (MLE) method and the method of moments. This paper will focus on the MLE for MGGDs. In [16], the scatter matrix is estimated based on the fixed-point (FP) iteration given a known $\beta \in (0, 1]$, which is proved to converge to the unique global optimal solution. If the shape parameter is unknown, a heuristic algorithm combining FP and Newton's method (a.k.a. Newton-Raphson method) was adopted. In [17], [18], it was proved that the MLE objective w.r.t. the scatter matrix is geodesic convex on the Riemannian manifold for all $\beta > 0$. And to tackle the cases when β is larger than 1, manifold-based methods like Riemannian averaged FP (RA-FP) algorithm [19] and Fisher scoring method [20] were proposed. The Riemannian averaged natural gradient method was also developed in [21] to make the parameter learning process amenable to the online scenarios.

In literature, most existing algorithms mainly discuss the estimation of the scatter matrix and the convergence is established when the shape parameter is known. When β is unknown, the convergence property of the heuristic alternating minimization based method is actually not evident [22]. As a matter of fact, the landscape of the negative log-likelihood function w.r.t. the shape parameter β is crucial to characterize the convergence of the existing estimation algorithms. In the literature, all the algorithms conduct the estimation of β by one-dimensional search algorithms such as the Newton's method. However, the local convergence nature of the Newton's method put the convergence guarantee of these algorithms at stake.

In this paper, to jointly learn all the model parameters we proposed two globally convergent algorithms based on the block majorization minimization (BMM) method [23], which has been studied in various application fields [24]–[26]. Besides that, we prove the negative log-likelihood func-

This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant 62001295 and in part by the Shanghai Sailing Program under Grant 20YF1430800.

[1]In the literature, $\boldsymbol{\Sigma}$ is commonly factorized as $\boldsymbol{\Sigma} = m\mathbf{M}$. It should be noted that algorithms proposed in this paper are also applicable to this case.

tion w.r.t. the shape parameter β is actually strictly convex for all $\beta > 0$, which to some extent reveals the underlying principle behind the success of many existing MGGD learning algorithms and to the best of our knowledge is the first result of this kind in the literature.

II. PROBLEM FORMULATION

Let $\mathbf{x}_i \in \mathbb{R}^p$ with $i = 1, 2, \ldots, N$ ($N \gg p$) follow an MGGD distribution with parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and β. The log-likelihood function for N independent and identically distributed samples $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ is accordingly given by

$$\begin{aligned}\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta) =& N \left[-\frac{p \log \pi}{2} + \log \Gamma(\frac{p}{2}) + \log \beta - \frac{p \log 2}{2} \frac{1}{\beta} \right. \\ & \left. - \log \Gamma(\frac{p}{2\beta}) \right] - \frac{N}{2} \log \det(\boldsymbol{\Sigma}) \\ & - \frac{1}{2} \sum_{i=1}^{N} \left((\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)^{\beta}.\end{aligned}$$

In this paper, we are interested in jointly learning all the MGGD parameters, i.e., $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta\}$ and we will focus on the problem when $\beta \in (0, 1]$ since values of β encountered in most practical applications belong to this interval [15], [16]. Based on the log-likelihood function above (ignoring the constants), the MLE problem for MGGD is as follows:

$$\begin{aligned}\underset{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \beta}{\text{minimize}} \; & - N \left[\log \beta - \frac{p \log 2}{2} \frac{1}{\beta} - \log \Gamma(\frac{p}{2\beta}) \right] \\ & + \frac{N}{2} \log \det(\boldsymbol{\Sigma}) + \frac{1}{2} \sum_{i=1}^{N} \left((\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)^{\beta} \quad (2) \\ \text{subject to} \; & \boldsymbol{\Sigma} \succeq \mathbf{0}, \; 0 \le \beta \le 1,\end{aligned}$$

which is a non-convex optimization problem[2]. In the literature, the location parameter $\boldsymbol{\mu}$ is always assumed to be known, say, by being estimated beforehand, and then we can always reduce the variable $\boldsymbol{\mu}$ from the log-likelihood function by centering each sample as $\overline{\mathbf{x}}_i = \mathbf{x}_i - \boldsymbol{\mu}$. In this paper, we will study jointly learning the model parameters, but the algorithm and convergence result are also applicable to the case of centered samples, i.e., $\boldsymbol{\mu} = \mathbf{0}$.

III. THE BMM METHOD

An optimization problem with the variable $\mathbf{x}$ partitioned into I blocks as $\mathbf{x} \triangleq (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^I)$ is given as follows:

$$\underset{\mathbf{x} \triangleq (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^I)}{\text{minimize}} \quad f\left(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^I\right) \quad \text{subject to} \; \mathbf{x}_i \in \mathcal{X}_i,$$

where $f : \mathbb{R}^N \to \mathbb{R}$ is a possibly non-convex objective function of $\mathbf{x} \in \mathbb{R}^N$ with $\mathbf{x}_i \in \mathbb{R}^{N_i}$ and $\sum_{i=1}^{I} N_i = N$, and $\mathcal{X}_i$'s are closed convex sets. Instead of dealing with the original optimization problem which could be difficult to tackle directly, starting from an initial point $\mathbf{x}_0$ the BMM method resolve the problem by solving a series of simple surrogate subproblems w.r.t. one single variable block each time [23]. Specifically, at the t-th iteration, the variable block $\mathbf{x}^i$ is updated according to the following update rules:

$$\begin{cases} \mathbf{x}_t^i \in \arg \min\limits_{\mathbf{x}_i \in \mathcal{X}_i} \overline{f_i}(\mathbf{x}^i, \mathbf{x}_{t-1}), \\ \mathbf{x}_t^{-i} = \mathbf{x}_{t-1}^{-i}, \text{ with } \mathbf{x}^{-i} \triangleq \left(\mathbf{x}^1, \ldots, \mathbf{x}^{i-1}, \mathbf{x}^{i+1}, \ldots, \mathbf{x}^I\right), \end{cases}$$

where $\overline{f_i}$ is a surrogate function majorizing f w.r.t. $\mathbf{x}^i$ [27]. The BMM method iteratively runs until some convergence criterion is met. The surrogate function in BMM can be chosen in a flexible way but a properly chosen one can lead to a light-weight iterative update while maintaining a fast convergence over iterations. In practice, the surrogate subproblems will be much applaudable if they are convex (hence, efficiently solvable) or obtain closed-form solutions.

IV. A TWO-BLOCK BMM ALGORITHM FOR MGGD PARAMETER LEARNING

In this section, we develop an efficient algorithm based on BMM to solve the MLE of MGGD given in Problem (2).

The $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$-Subproblem. For Problem (2), given the iterate $\{\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t\}$, the subproblem w.r.t. $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ is given by

$$\min_{\boldsymbol{\mu}, \boldsymbol{\Sigma} \succeq \mathbf{0}} \; \frac{N}{2} \log \det(\boldsymbol{\Sigma}) + \frac{1}{2} \sum_{i=1}^{N} \left((\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right)^{\beta_t},$$

where we denote the objective function as $F_{\mu,\Sigma}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t)$. Then we introduce the following useful result.

Lemma 1. *Function $f(y) = y^p$ ($0 \le p \le 1$), which is concave on $y \in [0, +\infty)$, is upperbounded at y_t as follows:*

$$y^p \le p y_t^{p-1} (y - y_t) + y_t^p.$$

Proof: It is easy to verify $f(y)$ is differentiable in y with a convex domain. It follows a function f is concave iff $f(y) \le f(x) + \nabla f(x)^T (y - x)$ for all $x, y \in \mathrm{dom} f$. ■

Based on Lemma 1, we can derive an upperbound function for $F_{\mu,\Sigma}(\boldsymbol{\mu}, \Sigma \mid \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t)$ at iterate $\{\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t\}$ as follows:

$$\begin{aligned}\overline{F_\mu}(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t) =& \frac{N}{2} \log \det(\boldsymbol{\Sigma}) \quad (3) \\ & + \frac{1}{2} \beta_t \sum_{i=1}^{N} \omega_i(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t) (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) + \text{const.},\end{aligned}$$

where $\omega_i(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t) = \left[(\mathbf{x}_i - \boldsymbol{\mu}_t)^T \boldsymbol{\Sigma}_t^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_t) \right]^{\beta_t - 1}$. Setting the partial derivatives of (3) w.r.t. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to zeros leads to the following update equations for $\{\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}\}$:

$$\begin{aligned}\boldsymbol{\mu}_{t+1} &= \frac{\sum_{i=1}^{N} \omega_i(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \mathbf{x}_i}{\sum_{i=1}^{N} \omega_i(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)} \\ \boldsymbol{\Sigma}_{t+1} &= \frac{\beta_t}{N} \sum_{i=1}^{N} \omega_i(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) (\mathbf{x}_i - \boldsymbol{\mu}_t)(\mathbf{x}_i - \boldsymbol{\mu}_t)^T. \end{aligned} \quad (4)$$

Lemma 2. *The pair $(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1})$ given by (4) uniquely minimizes the surrogate function in (3).*

Proof: This can be proved by solving the equivalent iterated minimization problem $\min_{\boldsymbol{\Sigma} \succeq \mathbf{0}} \min_{\boldsymbol{\mu}} F_{\mu,\Sigma}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. ■

The β-Subproblem. Given iterate $\{\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \beta_t\}$, the subproblem w.r.t. β is given in the following form:

[2]In the problem formulation, the constraints for $\boldsymbol{\Sigma}$ and β have been relaxed to be closed sets, which can be proved to be equivalent to the original ones by showing the objective is unbounded above at the boundary. The proof has been omitted in this conference paper.

Algorithm 1 A Two-Block BMM Alg. for MLE of MGGD

Initialization of $\{\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \beta_0\}$ and $t = 0$.
while $t \leq \text{MaxIteration}$ **do**
1) update $\{\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}\}$ by Equation (4)
2) update $\beta_{t+1} = \arg\min_{\beta\in[0,1]} F_\beta(\beta \mid \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \beta_t)$
3) $t \leftarrow t + 1$

end while

$$\min_{\beta\in[0,1]} \; -N\left[\log\beta - \frac{p\log 2}{2}\frac{1}{\beta} - \log\Gamma(\frac{p}{2\beta})\right] + \frac{1}{2}\sum_{i=1}^{N} y_{i,t}^{\beta},$$

where we have defined $y_{i,t} = (\mathbf{x}_i - \boldsymbol{\mu}_{t+1})^T \boldsymbol{\Sigma}_{t+1}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{t+1})$ and we also denote the objective as $F_\beta(\beta \mid \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \beta_t)$.

Theorem 3 (Strict Convexity of F_β)**.** *The objective function $F_\beta(\beta \mid \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \beta_t)$ with $\beta \in (0, +\infty)$ is strictly convex.*

Proof: Due to space limitation, we only provide a sketch of the proof. Some results on the digamma function $\psi(\cdot) = \frac{\Gamma'(\cdot)}{\Gamma(\cdot)}$ and trigamma function $\psi'(\cdot)$ are firstly given [28], [29].

Fact 1: As $x \to +\infty$, $\psi(x) \approx \log x - \frac{1}{2x} - \frac{1}{12x^2} + \frac{1}{120x^4} - \frac{1}{252x^6}$, and $\psi'(x) \approx \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} - \frac{1}{30x^5} + \frac{1}{42x^7}$.

Fact 2: When $x > 0$, $\psi(x+1) = \psi(x) + \frac{1}{x}$.

Fact 3: For all $x > 0$, $\psi(x) > \log\left(x + \frac{1}{2}\right) - \frac{1}{x}$.

Differentiating $F_\beta(\beta \mid \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \beta_t)$ twice will lead to $F''_\beta(\beta) = \frac{4N}{p^2} z^4 p(z) + q(z)$, where we have defined $z = \frac{p}{2\beta} > 0$, $p(z) = \frac{1}{z^2} + \frac{2}{z}(\log 2 + \psi(z)) + \psi'(z)$, and $q(z) = \frac{1}{2}\sum_{i=1}^{N} (\log y_i)^2 y_i^{\frac{p}{2z}}$ with $y_i = (\mathbf{x}_i - \boldsymbol{\mu}_{t+1})^T \boldsymbol{\Sigma}_{t+1}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}) > 0$. We will prove the strict convexity of the objective function $F_\beta(\beta)$ w.r.t. β by equivalently showing $F''_\beta(\beta)$ is strictly positive. From Fact 1, we have $\lim_{z\to+\infty} \frac{4}{p^2} z^4 p(z) > 0$. Thus, there exists a positive constant $N_0 > 1$ such that $p(z) > 0$ for all $z \geq N_0$. If we can prove $p(z) > p(z+1)$, we can get $p(z) > p(z+1) > 0$ for $z \in [N_0 - 1, N_0)$ and hence $p(z) > 0$ for all $z > 0$ by induction. Based on Fact 2, the problem that $p(z) > p(z+1)$ can be transformed to prove $\frac{(2\log 2 - 1)z^2 + (2\log 2 + 2)z + 2}{2z^2 + 2z} + \psi(z) > 0$. Further, with Fact 3 the above problem is reduced to proving $u(z) = \log 2 - \frac{z^2}{2z^2+2z} + \log\left(z + \frac{1}{2}\right) > 0$, which is straightforward by showing $\lim_{z\to 0^+} u(z) = 0$ and showing $u(z)$ is a monotonic decreasing function. Thus, we have $p(z) > 0$ for all $z > 0$. Combining that $q(z) > 0$ for all $z > 0$, we obtain that $F''_\beta(\beta) > 0$ for all $\beta > 0$. ∎

Based on Theorem 3, solution of the β-subproblem is unique. Many one-dimensional numerical methods can be used to solve for β, such as the bisection method or the Newton's method with line search. Finally, the overall BMM-based algorithm is summarized in Algorithm 1.

V. A THREE-BLOCK BMM ALGORITHM FOR MGGD PARAMETER LEARNING

In this section, we propose an alternative algorithm based on BMM where three variable blocks are involved.

The $\boldsymbol{\mu}$-Subproblem. Given the iterate $\{\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t\}$, the MLE subproblem w.r.t. variable $\boldsymbol{\mu}$ is given as follows:

$$\min_{\boldsymbol{\mu}} \quad \frac{1}{2}\sum_{i=1}^{N}\left((\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_t^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right)^{\beta_t}.$$

Algorithm 2 A Three-Block BMM Alg. for MLE of MGGD

Initialization of $\{\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \beta_0\}$ and $t = 0$.
while $t \leq \text{MaxIteration}$ **do**
1) update $\boldsymbol{\mu}_{t+1}$ by Equation (6)
2) update $\boldsymbol{\Sigma}_{t+1}$ by Equation (8)
3) update $\beta_{t+1} = \arg\min_{\beta} F_{\beta\in[0,1]}(\beta \mid \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \beta_t)$
4) $t \leftarrow t + 1$

end while

Denote the objective function as $F_\mu(\boldsymbol{\mu} \mid \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t)$, which is generally non-convex in $\boldsymbol{\mu}$ (when $\beta_t \neq 1$). Based on Lemma 1, we can obtain an upperbound function which is given by

$$\begin{aligned}&\overline{F_\mu}(\boldsymbol{\mu} \mid \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t) \qquad (5)\\ &= \frac{1}{2}\beta_t \sum_{i=1}^{N} \omega_i(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t)(\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}_t^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) + \text{const.},\end{aligned}$$

which becomes convex in $\boldsymbol{\mu}$. The $\boldsymbol{\mu}$-subproblem with the majorized objective (5) can be solved in closed-form by examining its first-order optimality condition given as follows:

$$\boldsymbol{\mu}_{t+1} = \frac{\sum_{i=1}^{N} \omega_i(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t)\mathbf{x}_i}{\sum_{i=1}^{N} \omega_i(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t)}. \qquad (6)$$

The $\boldsymbol{\Sigma}$-Subproblem. Given the iterate $\{\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_t, \beta_t\}$, the MLE subproblem w.r.t. $\boldsymbol{\Sigma}$ is accordingly given by

$$\min_{\boldsymbol{\Sigma}\succeq\mathbf{0}} \; \frac{N}{2}\log\det(\boldsymbol{\Sigma}) + \frac{1}{2}\sum_{i=1}^{N}\left((\mathbf{x}_i - \boldsymbol{\mu}_{t+1})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{t+1})\right)^{\beta_t}.$$

Denote the objective function as $F_\Sigma(\boldsymbol{\Sigma} \mid \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_t, \beta_t)$. Leveraging Lemma 1, we get the upperbound function as

$$\begin{aligned}&\overline{F_\Sigma}(\boldsymbol{\Sigma} \mid \boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_t, \beta_t) = \frac{N}{2}\log\det(\boldsymbol{\Sigma}) \qquad (7)\\ &+\frac{1}{2}\beta_t \sum_{i=1}^{N} \omega_i(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_t, \beta_t)(\mathbf{x}_i - \boldsymbol{\mu}_{t+1})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_{t+1}) + \text{const.},\end{aligned}$$

which is convex in $\boldsymbol{\Sigma}^{-1}$. A unique and optimal solution can be attained for the update of $\boldsymbol{\Sigma}_{t+1}$ which is given by

$$\boldsymbol{\Sigma}_{t+1} = \frac{\beta_t}{N}\sum_{i=1}^{N} \omega_i(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_t, \beta_t)(\mathbf{x}_i - \boldsymbol{\mu}_{t+1})(\mathbf{x}_i - \boldsymbol{\mu}_{t+1})^T. \qquad (8)$$

It should be noted that, Equation (8) resembles the FP update proposed in [16]. Therefore, the FP algorithm in [16] can be interpreted from a majorization minimization perspective.

The β-Subproblem. Given the iterate $\{\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}, \beta_t\}$, the resolution of the β-subproblem is the same as that in the last section and hence can be solved accordingly. Finally, the overall BMM algorithm is summarized in Algorithm 2.

Proposition 4 (Convergence Property)**.** *Suppose the MLE estimates of Problem (2) exist, then every limit point, denoted by $\{\boldsymbol{\mu}_\infty, \boldsymbol{\Sigma}_\infty, \beta_\infty\}$, of the sequence $\{\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \beta_t\}$ generated by the BMM-based algorithms (i.e., Algorithm 1 and Algorithm 2) is a stationary point of Problem (2).*

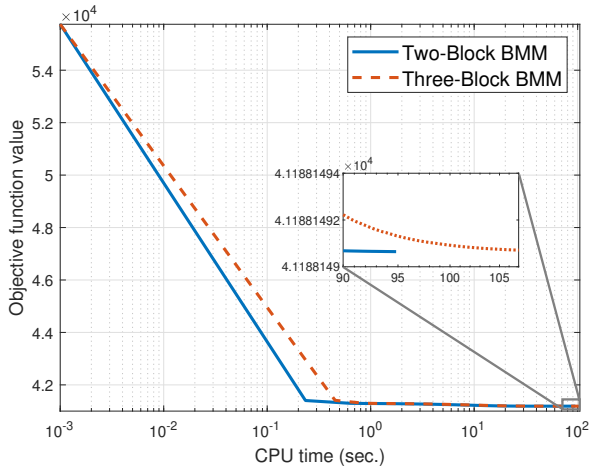

Fig. 1. Convergence comparisons between two proposed BMM algorithms ($p = 3, n = 10000, \beta = 0.8, \boldsymbol{\mu} = \mathbf{1}$).

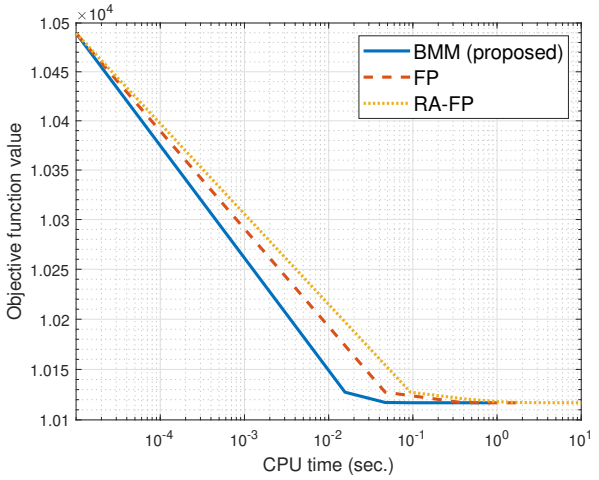

Fig. 2. Convergence comparisons between the proposed BMM algorithm and existing algorithms ($p = 4, n = 1000, \beta = 0.5$).

VI. NUMERICAL SIMULATIONS

In this section, numerical simulations are conducted to demonstrate the performance of the proposed BMM-based algorithms with comparisons to some state-of-the-art algorithms. The synthetic data is used where $\mathbf{x} \in \mathbb{R}^p$ is generated based on the following stochastic representation [2]:

$$\mathbf{x} = \tau \boldsymbol{\Sigma}^{\frac{1}{2}} \mathbf{u} + \boldsymbol{\mu},$$

where τ is a scalar random variable and $\boldsymbol{\Sigma}$ is the scatter matrix which is specified as

$$\boldsymbol{\Sigma}(i,j) = m\gamma^{|i-j|}, \;\; i,j = \{1,2,...,p\},$$

where we have set $m = 2$ and $\gamma = 0.5$.

We first compare the convergence property of the two-block BMM-based algorithm and three-block BMM-based algorithm. In BMM, the bisection method has been applied to determine the shape parameter β. The convergence criterion for all the implemented algorithms in this paper is set as $\frac{\|\boldsymbol{\mu}_{t+1}-\boldsymbol{\mu}_t\|_2^2}{\|\boldsymbol{\mu}_t\|_2^2} < 10^{-5}$, $\frac{\|\boldsymbol{\Sigma}_{t+1}-\boldsymbol{\Sigma}_t\|_F^2}{\|\boldsymbol{\Sigma}_t\|_F^2} < 10^{-5}$, and $\frac{|\beta_{t+1}-\beta_t|}{|\beta_t|} < 10^{-5}$. As shown in Fig. 1, we have observed that these two algorithms can converge to the same objective value and the two-block BMM generally obtains a faster convergence rate compared to the three-block counterpart. We further demonstrate the convergence comparisons among BMM, FP [16], and RA-FP [19] which is given in Fig. 2. Since $\boldsymbol{\mu}$ is not estimated in FP and RA-FP, we have ignored the estimation of $\boldsymbol{\mu}$ in our BMM-based algorithm. It can be observed that BMM outperforms the other two algorithms in terms of CPU times.

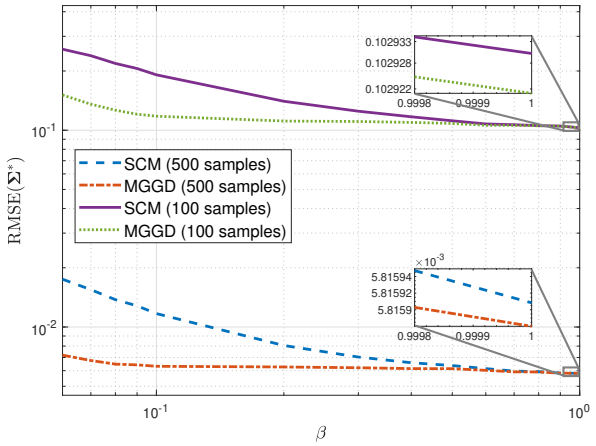

Fig. 3. Estimation performance of scatter matrix ($p = 10, \boldsymbol{\mu} = \mathbf{0}$).

Then, we demonstrate the estimation accuracy of the scatter matrices. In Fig. 3, we compared our BMM-based method for MGGD MLE with the sample covariance matrix (SCM), which is equivalent to Gaussian MLE. Since the scatter matrix is a constant-scaled version of the covariance matrix for MGGD [30], both the estimated values from different methods and $\boldsymbol{\Sigma}_{\text{true}}$ are normalized by their trace values. All the experimental results are averaged over 1000 Monte Carlo simulations and the estimation error is measured by the relative mean squared error (RMSE) defined as $\text{RMSE}(\boldsymbol{\Sigma}^\star) = \frac{\|\boldsymbol{\Sigma}^\star - \boldsymbol{\Sigma}_{\text{true}}\|_F^2}{\|\boldsymbol{\Sigma}_{\text{true}}\|_F^2}$. In Fig. 3, as expected the results from MGGD can outperform the sample covariance in terms of estimation accuracy. The superior performance of MGGD over sample covariance is more pronounced when β is smaller. Also, with the increasing of sample size N, the estimation errors of all algorithms decrease accordingly.

VII. CONCLUSIONS

In this paper, two globally convergent algorithms based on the block majorization minimization method have been proposed to jointly learn all the model parameters based on the maximum likelihood estimation, which to our best knowledge is the first convergent algorithm of this type. Moreover, the objective function with respect to the shape parameter is proved to be strictly convex, which to some extent explains the underlying principle behind the success of many existing MGGD learning algorithms.

978-1-7281-5767-2/21/$31.00 ©2021 IEEE

VIII. REFERENCES

[1] I. R. Goodman and S. Kotz, "Multivariate θ-generalized normal distributions," *Journal of Multivariate Analysis*, vol. 3, no. 2, pp. 204–219, 1973.

[2] E. Gómez, M. Gomez-Viilegas, and J. M. Marín, "A multivariate generalization of the power exponential family of distributions," *Communications in Statistics-Theory and Methods*, vol. 27, no. 3, pp. 589–600, 1998.

[3] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 909–919, 1999.

[4] M. S. Allili, D. Ziou, N. Bouguila, and S. Boutemedjet, "Image and video segmentation by combining unsupervised generalized Gaussian mixture modeling and feature selection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 10, pp. 1373–1377, 2010.

[5] M. Bicego, D. Gonzalez-Jimenez, E. Grosso, and J. A. Castro, "Generalized Gaussian distributions for sequential data classification," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[6] H. Hristova, O. Le Meur, R. Cozot, and K. Bouatouch, "Transformation of the multivariate generalized Gaussian distribution for image editing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 10, pp. 2813–2826, 2017.

[7] Q. Z. Ahmed, K.-H. Park, and M.-S. Alouini, "Ultrawide bandwidth receiver based on a multivariate generalized Gaussian distribution," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 1800–1810, 2014.

[8] S. Le Cam, A. Belghith, C. Collet, and F. Salzenstein, "Wheezing sounds detection using multivariate generalized Gaussian distributions," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 541–544.

[9] M. N. Desai and R. S. Mangoubi, "Robust Gaussian and non-Gaussian matched subspace detection," *IEEE Transactions on Signal Processing*, vol. 51, no. 12, pp. 3115–3127, 2003.

[10] X. Jiang, W.-J. Zeng, A. Yasotharan, H. C. So, and T. Kirubarajan, "Minimum dispersion beamforming for non-Gaussian signals," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1879–1893, 2014.

[11] Z. Zhao and D. P. Palomar, "Robust maximum likelihood estimation of sparse vector error correction model," in *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2017, pp. 913–917.

[12] Z. Zhao, R. Zhou, and D. P. Palomar, "Optimal mean-reverting portfolio with leverage constraint for statistical arbitrage in finance," *IEEE Transactions on Signal Processing*, vol. 67, no. 7, pp. 1681–1695, 2019.

[13] S. Kotz, "Multivariate distributions at a cross road," in *A Modern Course on Statistical Distributions in Scientific Work*. Springer, 1975, pp. 247–270.

[14] A. Dytso, R. Bustin, H. V. Poor, and S. Shamai, "Analytical properties of generalized Gaussian distributions," *Journal of Statistical Distributions and Applications*, vol. 5, no. 1, pp. 1–40, 2018.

[15] J. R. Hernandez, M. Amado, and F. Perez-Gonzalez, "DCT-domain watermarking techniques for still images: Detector performance analysis and a new structure," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 55–68, 2000.

[16] F. Pascal, L. Bombrun, J.-Y. Tourneret, and Y. Berthoumieu, "Parameter estimation for multivariate generalized Gaussian distributions," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 5960–5971, 2013.

[17] S. Sra and R. Hosseini, "Geometric optimisation on positive definite matrices for elliptically contoured distributions," *Advances in Neural Information Processing Systems*, vol. 26, pp. 2562–2570, 2013.

[18] T. Zhang, A. Wiesel, and M. S. Greco, "Multivariate generalized Gaussian distribution: Convexity and graphical models," *IEEE Transactions on Signal Processing*, vol. 61, no. 16, pp. 4141–4148, 2013.

[19] Z. Boukouvalas, S. Said, L. Bombrun, Y. Berthoumieu, and T. Adalı, "A new Riemannian averaged fixed-point algorithm for MGGD parameter estimation," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2314–2318, 2015.

[20] Z. Boukouvalas, G.-S. Fu, and T. Adalı, "An efficient multivariate generalized Gaussian distribution estimator: Application to IVA," in *2015 49th Annual Conference on Information Sciences and Systems*. IEEE, 2015, pp. 1–4.

[21] J. Zhou, S. Said, and Y. Berthoumieu, "Online estimation of MGGD: the Riemannian Averaged Natural Gradient method," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*. IEEE, 2019, pp. 515–519.

[22] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1995.

[23] M. Razaviyayn, M. Hong, and Z.-Q. 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] Z. Zhao and D. P. Palomar, "Mean-reverting portfolio with budget constraint," *IEEE Transactions on Signal Processing*, vol. 66, no. 9, pp. 2342–2357, 2018.

[25] ——, "Sparse reduced rank regression with nonconvex regularization," in *2018 IEEE Statistical Signal Processing Workshop (SSP)*. IEEE, 2018, pp. 811–815.

[26] ——, "MIMO transmit beampattern matching under waveform constraints," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 3281–3285.

[27] 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.

[28] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 9th printing. Dover, 1972.

[29] N. Elezovic, C. Giordano, and J. Pecaric, "The best bounds in Gautschi's inequality," *Mathematical Inequalities & Applications*, vol. 3, no. 2, pp. 239–252, 2000.

[30] E. Ollila, D. E. Tyler, V. Koivunen, and H. V. Poor, "Complex elliptically symmetric distributions: Survey, new results and applications," *IEEE Transactions on signal processing*, vol. 60, no. 11, pp. 5597–5625, 2012.

978-1-7281-5767-2/21/$31.00 ©2021 IEEE