



Riemannian information gradient methods for the parameter estimation of ECD : Some applications in image processing

Jialun Zhou, Salem Said, Yannick Berthoumieu

► To cite this version:

Jialun Zhou, Salem Said, Yannick Berthoumieu. Riemannian information gradient methods for the parameter estimation of ECD : Some applications in image processing. Signal Processing, 2022, 10.1016/j.sigpro.2021.108376 . hal-03866181

HAL Id: hal-03866181

<https://hal.science/hal-03866181>

Submitted on 22 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Riemannian information gradient methods for the parameter estimation of ECD : Some applications in image processing

Jialun Zhou, Salem Said, Yannick Berthoumieu^a

^a *University Bordeaux, CNRS, IMS, UMR 5218, Groupe Signal et Image, F-33405 Talence, France*

Abstract

Elliptically-contoured distributions (ECD) play a significant role, in computer vision, image processing, radar, and biomedical signal processing. Maximum likelihood estimation (MLE) of ECD leads to a system of non-linear equations, most-often addressed using fixed-point (FP) methods. Unfortunately, the computation time required for these methods is unacceptably long, for large-scale or high-dimensional datasets. To overcome this difficulty, the present work introduces a Riemannian optimisation method, the information stochastic gradient (ISG). The ISG is an online (recursive) method, which achieves the same performance as MLE, for large-scale datasets, while requiring modest memory and time resources. To develop the ISG method, the Riemannian information gradient is derived taking into account the product manifold associated to the underlying parameter space of the ECD. From this information gradient definition, we define also, the information deterministic gradient (IDG), an offline (batch) method, which is an alternative, for moderate-sized datasets. The present work formulates these two methods, and demonstrates their performance through numerical simulations. Two

applications, to image re-colorization, and to texture classification, are also worked out.

Keywords: elliptically-contoured distribution, Riemannian information gradient, large-scale dataset, image re-colorization, texture classification.

1. Introduction

The family of Elliptically-contoured distributions (ECD) was originally introduced in [1], and investigated in [2, 3]. It contains many widely-used statistical distributions, such as elliptical Gamma, Pearson type II, and elliptical multivariate logistic distributions. In terms of applications, the most popular classes of ECD are multivariate generalized Gaussian distributions (MGGD), and multivariate Student-T distributions [4, 5, 6]. These are location-scale distributions, and are further parameterised by a shape parameter, or a degrees of freedom parameter.

MGGD are used in image processing, as models for wavelet and curvelet coefficients, and as models for three-channel color vectors, in image denoising, context-based image retrieval, image thresholding, texture classification, and image quality assessment [7, 8, 9, 10, 11, 12]. MGGD are also used in video coding and denoising, radar signal processing, and biomedical signal processing [13, 14].

Some applications of Student-T distributions are presented in [15], involving image denoising. In radar imaging, the Student-T distribution, so-called \mathcal{G}^0 model within the family of spherically invariant random vectors (SIRVs), is largely exploited in the context of SAR or PolSAR imaging, for tasks such as despeckling, classification, segmentation or detection [16, 17, 18].

Because ECD have been successful in real-world image and signal processing applications, much attention has been devoted to developing efficient methods for estimating their parameters. The vast majority of works, dedicated to this estimation problem, are focused on estimating the scatter matrix, considering the other parameters, i.e. location and shape parameters, as known.

In terms of maximum-likelihood estimation, two main classes of algorithms have been studied. Fixed-point (FP) algorithms, and gradient descent algorithms have been proposed, based on the geometric properties of the manifold of positive definite matrices [19, 20, 21, 22, 23]. For MGGD, when the location parameter is equal to zero and the shape parameter is given, the uniqueness of the maximum-likelihood estimator has been shown, under a restriction on the value of the shape parameter [24]. In this case, a method of moments has also been developed [25]. For Student-T distributions, with a known degrees of freedom parameter, a fixed-point method for parameter estimation is given in [15], where the existence and uniqueness, of location and scatter maximum-likelihood estimates, is shown for a fixed degrees of freedom parameter, superior to 1.

There is a shared drawback, in all of the maximum-likelihood estimation methods, just mentioned [21, 22, 15, 25, 24]. Specifically, these methods work well for datasets of moderate size and dimension, but require excessive resources in memory and time, for large-scale datasets, e.g. with a few millions of samples, an order of magnitude commonly encountered in optical or SAR image processing. This issue can be so severe as to make any of these methods inapplicable.

Mainly, this is due to the fact that all of these methods are off-line, or batch, estimation methods. They require access to the whole dataset, at once, for each iteration, and therefore consume increasing time and memory resources, in order to converge to a useful estimate, as the dataset grows large.

In order to overcome this drawback, the present work builds on the ideas from Riemannian stochastic optimisation, proposed in [26, 27]. The problem of estimating the parameters of an ECD is viewed as the problem of minimising the Kullback-Leibler divergence, between the true (unknown) and estimate distributions. When this problem is addressed using Riemannian stochastic optimisation, each iteration of a stochastic optimisation method requires access to only one datapoint (one sample), instead of the whole dataset. In this way, the present work proposes a recursive method, for estimating the parameters of ECD (each time a new sample is processed, this sample is used to update the current estimate).

The proposed method will be called the information stochastic gradient (ISG). In its simplest form, it is an improvement of a previous method, used to estimate the scatter matrix, when the location and shape parameters are known [28]. In this paper, we consider also the general case where the scatter matrix, the location and shape parameters are unknown. The ISG method relies on two main ideas :

- The greatest difficulty, in using recursive methods, is that they may require a careful choice of step-sizes. The standard Riemannian stochastic gradient method (as in [26]), is very sensitive to the choice of step-sizes. However, using the information gradient (also called the natural gradi-

ent [29, 30]) leads to an automatic choice of step-sizes, which guarantees optimal performance. The ISG method implements the information gradient, relying on the Fisher information metric (or matrix), of the ECD model.

- The parameter space of an ECD model does not only contain the scatter matrix, but also location and shape parameters. In the case of MGGD or Student-T models, this parameter space is a product space, made up of triplets: (scatter matrix, location parameter, shape/degrees of freedom parameter). Since the geodesic curves of this space do not have a tractable expression [31], an intuitive idea is to update each one of the three parameters, in its own turn, in an alternating fashion.

To understand the benefit of combining these two ideas, consider the special case of MGGD. In this case, a method of moments (MM) was used for the joint estimation of all three parameters [25], while their maximum-likelihood estimation (MLE) was studied in [24]. It is well known that MLE performs better than MM in most scenarios [32]. However, as mentioned above, MLE cannot be applied to large scale datasets, due to its computational requirements. The ISG method strikes a balance between the low complexity of MM, and the stronger performance of MLE. For example, in the case where the scatter matrix and the location parameter are unknown, the complexity of ISG is comparable to that of the MM, while its performance is similar to that of the MLE, when the number of available samples is sufficiently large. In other words, the size of the dataset is leveraged as a source of information, rather than as a computational burden.

The two ideas which underly the ISG method (discussed above), are also

implemented in an offline (batch) method, called the information deterministic gradient (IDG) method. While its complexity (and therefore computation time) is much higher than the ISG method, the IDG method consistently outperforms other methods, even in the case where all three parameters are unknown.

The main results of the present work are given in Section 4. A detailed comparison of various estimation methods (MM, FP, ISG, IDG), based on computer experiments, is carried out in Section 5. Two image processing applications with real datasets are presented in Section 6.

First, Sections 2 and 3 define the general estimation problem for ECD models, and introduce some necessary geometric concepts.

2. The estimation problem

2.1. The ECD family

ECD is a general family of probability distributions that contains many important sub-families. The name ECD comes from the fact that when an ECD has a probability density function, the contours (level surfaces) of this function are ellipsoids.

The location, or expectation, parameter μ of an ECD determines the centre of these ellipsoids, while the axes of these ellipsoids are proportional to the eigenvalues of the inverse of the scatter matrix Σ . The shape parameter β determines the factor for this proportionality (β is the degrees of freedom parameter, for Student-T distributions).

Let X be a m -dimensional random vector that follows a ECD model. Denote $\theta = (\mu, \Sigma, \beta)$ the parameters of this ECD, and $\Theta = \mathbb{R}^m \times \mathcal{P}_m \times \mathbb{R}_+$

its parametric space, where \mathcal{P}_m is the set of all symmetric positive definite matrices of size $m \times m$. If X has a probability density function, then this takes on the following form

$$p(x; \theta) = c(\beta) |\Sigma|^{-1/2} g[\delta_x(\mu, \Sigma), \beta] \quad (1)$$

where $c(\beta)$ is a normalizing factor which depends only on β , and $\delta_x(\mu, \Sigma) = (x - \mu)^\dagger \Sigma^{-1} (x - \mu)$. The density generator g depends on the specific sub-family of ECD distributions, for example

$$\begin{aligned} g[\delta_x(\mu, \Sigma), \beta] &= \exp\left(-\frac{1}{2}\delta_x^\beta\right) \text{ for MGGD} \\ g[\delta_x(\mu, \Sigma), \beta] &= \left(1 + \frac{1}{\beta}\delta_x\right)^{-\frac{\beta+m}{2}} \text{ for Student-T} \end{aligned}$$

2.2. Problem formulation

Parameter estimation will be formulated as the problem of minimising the Kullback-Leibler divergence $D(\theta^* || \theta)$, denoted $D(\theta)$ for short. That is to say, the estimator $\hat{\theta}$ is sought which is the solution of the following minimisation problem

$$\hat{\theta} = \arg \min_{\theta \in \Theta} D(\theta) \quad (2)$$

Recall the definition of the KL divergence

$$\begin{aligned} D(\theta) &= \int_{\mathbb{R}^p} p(x; \theta^*) \ln \left(\frac{p(x; \theta^*)}{p(x; \theta)} \right) dx \\ &= \mathbb{E}_{\theta^*} [\ell(\theta^*; x)] - \mathbb{E}_{\theta^*} [\ell(\theta; x)] \end{aligned} \quad (3)$$

Where $\ell(\theta; x) = \log p(x; \theta)$ is the log-likelihood,

$$\ell(\theta; x) = \alpha(\beta) - \frac{1}{2} \log \det(\Sigma) + h(\delta_x, \beta) \quad (4)$$

with $\alpha(\beta) = \log c(\beta)$ and $h = \log g$. In the following, the KL divergence (3) will be minimised using Riemannian information gradient descent. Some Riemannian Information geometry concepts are recalled in the following section.

3. Necessary geometric concepts

The gradient descent method on Riemannian manifolds is based on the following update rule [33]

$$\theta_{n+1} = R_{\theta_n}(\alpha_{n+1}u(\theta_n)) \quad (5)$$

Here, the smooth mapping R_θ from the tangent space $T_\theta\Theta$ to Θ is required to be a retraction, in the sense that it verifies

$$R_\theta(0_\theta) = \theta \quad (6a)$$

$$DR_\theta(0_\theta) = \text{Id}_{T_\theta\Theta} \quad (6b)$$

where 0_θ denotes the zero element in $T_\theta\Theta$, and $\text{Id}_{T_\theta\Theta}$ denotes the identity mapping on $T_\theta\Theta$. Each vector $u(\theta_n)$ belongs to the tangent space $T_{\theta_n}\Theta$, and provides the direction of descent. In the present work, $-u(\theta_n)$ is the Riemannian information gradient, derived using the Fisher information metric. The positive scalar α_n is the step-size. The aim of equation (5) is to generate a sequence $(\theta_n)_{n \geq 0} \in \Theta$ that converges to a stationary point θ^* of the cost function (under some restrictions on u and α).

For our estimation problem, the model has three different parameters, which belong to three different Riemannian manifolds. Precisely, the parameter space is the product manifold $\Theta = \mathbb{R}^m \times \mathcal{P}_m \times \mathbb{R}_+$. Therefore, tractable

expressions of the Fisher information metric, and of the intrinsic geodesic map, on this product manifold, are needed. However, Θ does not support any such expressions [31]. As the global Fisher information metric has not a closed form, we propose to use the product metric

$$\langle u, v \rangle_\theta = \langle u_\mu, v_\mu \rangle_\mu + \langle u_\Sigma, v_\Sigma \rangle_\Sigma + \langle u_\beta, v_\beta \rangle_\beta \quad (7)$$

where $u = (u_\mu, u_\Sigma, u_\beta)$ and $v = (v_\mu, v_\Sigma, v_\beta)$ are tangent vectors at the point $\theta = (\mu, \Sigma, \beta)$. The metrics $\langle \cdot, \cdot \rangle_\mu$, $\langle \cdot, \cdot \rangle_\Sigma$ and $\langle \cdot, \cdot \rangle_\beta$ are respectively the intrinsic Fisher information metrics of their corresponding sub-spaces. For the location parameter μ in \mathbb{R}^m , its information metric is expressed in terms of the usual Euclidean metric,

$$\langle u_\mu, v_\mu \rangle_\mu = I_\mu u_\mu^\dagger \Sigma^{-1} v_\mu \quad (8)$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product in Euclidean space. The information constant I_μ is

$$I_\mu = -\frac{4}{m} \mathbb{E} \left[\frac{\partial^2 h(\delta_x, \beta)}{\partial \delta_x^2} \delta_x \right] - 2 \mathbb{E} \left[\frac{\partial h(\delta_x, \beta)}{\partial \delta_x} \right] \quad (9)$$

with $h = \log g$. As for $\Sigma \in \mathcal{P}_m$, the Fisher information metric $\langle \cdot, \cdot \rangle_\Sigma$ for the ECD model is defined by the Riemannian geometry of \mathcal{P}_m [34].

$$\langle U_\Sigma, V_\Sigma \rangle_\Sigma = I_1 \text{tr}(\Sigma^{-1} U_\Sigma \Sigma^{-1} V_\Sigma) + I_2 \text{tr}(\Sigma^{-1} U_\Sigma) \text{tr}(\Sigma^{-1} V_\Sigma) \quad (10)$$

Here the constants $I_1 > 0$ and $I_2 \geq 0$ depend on the particular model under consideration, as follows

$$I_1 = \frac{2\mathcal{A}}{m(m+2)} \quad I_2 = \frac{\mathcal{A}}{m(m+2)} - \frac{1}{4} \quad (11)$$

$$\mathcal{A} = \mathbb{E} \left[\left(\frac{\partial h(\delta_x, \beta)}{\partial \delta_x} \delta_x \right)^2 \right]$$

The shape factor β belongs to \mathbb{R}_+ , so the Fisher information metric is given by

$$\langle u_\beta, v_\beta \rangle_\beta = I_\beta u_\beta v_\beta \quad (12)$$

with the information constant

$$I_\beta = -\mathbb{E} \left[\frac{\partial^2 \alpha(\beta)}{\partial \beta^2} + \frac{\partial^2 h(\delta_x, \beta)}{\partial \beta^2} \right] \quad (13)$$

Now, the information gradient $\nabla_\theta D(\theta)$ with respect to the product metric (7) is obtained by solving the following equation,

$$d D(\theta) v = \langle \nabla_\theta D(\theta), v \rangle_\theta \quad (14)$$

where the scalar product on the right-hand side is given by (7), and d is the differential form of D . Precisely, this product information gradient has the following form

$$\nabla_\theta D(\theta) = (\nabla_\mu D(\theta), \nabla_\Sigma D(\theta), \nabla_\beta D(\theta)) \quad (15)$$

The first component $\nabla_\mu D(\theta)$ is expressed as

$$\nabla_\mu D(\theta) = -I_\mu^{-1} \Sigma \mathbb{E} [G_\mu(\theta; x)] \quad (16)$$

where I_μ is given in equation (9), and vector $G_\mu(\theta; x)$ is actually the gradient in the classic Euclidean sense

$$G_\mu(\theta; x) = 2 \frac{\partial h(\delta_x, \beta)}{\partial \delta_x} \Sigma^{-1} (x - \mu) \quad (17)$$

The second component $\nabla_\Sigma D(\theta)$ is a bit more complicated (see Figure 3, for an illustration of the following computations)

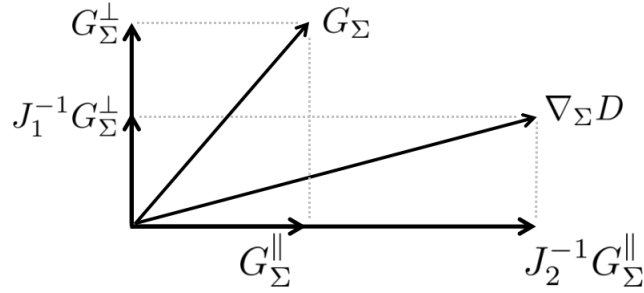
$$\nabla_\Sigma D(\theta) = -J_1^{-1} \mathbb{E} \left\{ [G_\Sigma(\Sigma; x)]^\perp \right\} - J_2^{-1} \mathbb{E} \left\{ [G_\Sigma(\theta; x)]^\parallel \right\} \quad (18)$$

where $J_1 = I_1$ and $J_2 = I_1 + mI_2$, in terms of I_1 and I_2 given in (11), and where \perp and \parallel denote the following decomposition of $G_\Sigma(\theta; x)$,

$$[G_\Sigma(\theta; x)]^\parallel = \frac{1}{m} \text{tr} [\Sigma^{-1} G_\Sigma(\theta; x)] \Sigma \quad (19)$$

$$[G_\Sigma(\theta; x)]^\perp = G_\Sigma(\theta; x) - [G_\Sigma(\theta; x)]^\parallel \quad (20)$$

in terms of



$$G_\Sigma(\theta; x) = -\frac{1}{2} \Sigma - \frac{\partial h(\delta_x, \beta)}{\partial \delta_x} S_x \quad \text{with} \quad S_x = (x - \mu)(x - \mu)^\dagger \quad (21)$$

Finally, for the third component,

$$\nabla_\beta D(\theta) = -I_\beta^{-1} \mathbb{E} [G_\beta(\beta; x)] \quad (22)$$

where I_β was given in (13), and

$$G_\beta(\theta; x) = \frac{\partial \alpha(\beta)}{\partial \beta} - \frac{\partial h(\delta_x, \beta)}{\partial \beta} \quad (23)$$

With regard to the retraction R_θ , it will be defined as the product Riemannian exponential map,

$$\begin{aligned} R_\theta : \quad T_\theta \Theta &\longrightarrow \Theta \\ u = \begin{pmatrix} u_\mu \\ u_\Sigma \\ u_\beta \end{pmatrix} &\longmapsto \begin{pmatrix} \text{Exp}_\mu(u_\mu) \\ \text{Exp}_\Sigma(u_\Sigma) \\ \text{Exp}_\beta(u_\beta) \end{pmatrix} \end{aligned} \quad (24)$$

where u is the direction of descent. The exponential map on \mathbb{R}^p (a Euclidean space) reduces to vector addition

$$\text{Exp}_\mu(u_\mu) = \mu + u_\mu \quad (25)$$

The exponential map on \mathcal{P}_m is defined as follows [35]:

$$\text{Exp}_\Sigma(u_\Sigma) = \Sigma \exp(\Sigma^{-1}u_\Sigma) \quad (26)$$

As for β , since it belongs to \mathbb{R}_+ , the corresponding exponential map is a 1-dimensional version of (26)

$$\text{Exp}_\beta(v_\beta) = \beta \exp(\beta^{-1}v_\beta) \quad (27)$$

All these three exponential map Exp verify the properties (6), therefore the their direct product (24) also verifies these properties, and is a well-defined retraction. Finally, the Riemannian distance associated to the metric (7) is given by

$$d^2(\theta_1, \theta_2) = d_{\mathbb{R}^m}^2(\mu_1, \mu_2) + d_{\mathcal{P}_m}^2(\Sigma_1, \Sigma_2) + d_{\mathbb{R}_+}^2(\beta_1, \beta_2) \quad (28)$$

For μ , the information distance is proportional to the Euclidean distance in \mathbb{R}^p

$$d_{\mathbb{R}^m}^2(\mu_1, \mu_2) = I_\mu (\mu_1 - \mu_2)^\dagger (\mu_1 - \mu_2) \quad (29a)$$

for $\mu_1, \mu_2 \in \mathbb{R}^m$, with the constant I_μ given by (9). For Σ , the information distance is defined as in [36]

$$d_{\mathcal{P}_m}^2(\Sigma_1, \Sigma_2) = I_1 \text{tr} [\log(\Sigma_1^{-1} \Sigma_2)]^2 + I_2 \text{tr}^2 [\log(\Sigma_1^{-1} \Sigma_2)] \quad (29b)$$

for $\Sigma_1, \Sigma_2 \in \mathcal{P}_m$, where the constants I_1 and I_2 are given by (11), and the function \log denotes the symmetric matrix logarithm. Finally, for β , the

information distance is given by

$$d_{\mathbb{R}_+}^2(\beta_1, \beta_2) = I_\beta \log^2(\beta_1^{-1} \beta_2) \quad (29c)$$

for $\beta_1, \beta_2 \in \mathbb{R}_+$, where I_β is given in (13). With the necessary geometric concepts now in place, the next section will introduce our estimation algorithms.

4. The IDG and ISG methods

This section will describe the IDG and ISG methods, and discuss their main properties. The IDG method (information deterministic gradient) is a deterministic gradient method, and the ISG method (information stochastic gradient) is a stochastic gradient method.

When the direction of descent is chosen according to (15), the updated estimates $\theta_{k+1} = (\mu_{k+1}, \Sigma_{k+1}, \beta_{k+1})$ rely on the current estimates $\theta_k = (\mu_k, \Sigma_k, \beta_k)$, through the following alternating optimisation scheme

$$\begin{aligned} \text{step 1 : } \mu_{k+1} &\leftarrow (\mu_k, \Sigma_k, \beta_k) \\ \text{step 2 : } \Sigma_{k+1} &\leftarrow (\mu_{k+1}, \Sigma_k, \beta_k) \\ \text{step 3 : } \beta_{k+1} &\leftarrow (\mu_{k+1}, \Sigma_{k+1}, \beta_k) \end{aligned} \quad (30)$$

4.1. Deterministic gradient

The IDG method is a second-order offline method, somewhat similar to a Newton method. In the Newton method, the direction of descent is found by solving the Newton equation [33]. In the IDG method, the Hessian in the Newton equation is approximated by the Fisher information metric (or matrix) $\mathcal{I}(\theta)$.

Since IDG is an offline method, it chooses a direction of descent which depends on the complete dataset. The cost function (2) is reformulated, by

replacing the KL divergence, with the empirical average of $-\ell(\theta, x_n)$. This empirical average is denoted by $\hat{D}(\theta)$,

$$\hat{D}(\theta) = -\frac{1}{N} \sum_{n=1}^N \ell(\theta_n; x_n) \quad (31)$$

If the current estimate is θ_k , the direction of descent is given by its three components

$$\nabla_{\mu} \hat{D}(\theta_k) = -I_{\mu}^{-1} \Sigma \frac{1}{N} \sum_{n=1}^N G_{\mu}(\theta_k; x_n) \quad (32a)$$

$$\nabla_{\Sigma} \hat{D}(\theta_k) = -J_1^{-1} \frac{1}{N} \sum_{n=1}^N [G_{\Sigma}(\theta_k; x_n)]^{\perp} - J_2^{-1} \frac{1}{N} \sum_{n=1}^N [G_{\Sigma}(\theta_k; x_n)]^{\parallel} \quad (32b)$$

$$\nabla_{\beta} \hat{D}(\theta_k) = -I_{\beta}^{-1} \frac{1}{N} \sum_{n=1}^N G_{\beta}(\theta_k; x_n) \quad (32c)$$

which are the same as (16), (18), (22), but with expectations replaced by empirical averages. Using the expressions (32a), (32b), (32c), the IDG algorithm can now be stated as follows.

Algorithm 1 Information Deterministic Gradient (IDG) algorithm

Input: A dataset $\mathcal{X} = (x_1, \dots, x_N)$, an initialization $\theta_0 \in S_0 \subset \Theta$;

Output: The estimate $\hat{\theta}$;

1: **for** $k = 0, 1, 2, \dots, K$ **do**

2: $\mu_{k+1} \leftarrow \mu_k - \alpha_{\mu} \nabla_{\mu} \hat{D}(\mu_k, \Sigma_k, \beta_k);$

3: $\Sigma_{k+1} \leftarrow \Sigma_k \exp \left(-\Sigma_k^{-1} \alpha_{\Sigma} \nabla_{\Sigma} \hat{D}(\mu_{k+1}, \Sigma_k, \beta_k) \right);$

4: $\beta_{k+1} \leftarrow \beta_k \exp \left(-\beta_k^{-1} \alpha_{\beta} \nabla_{\beta} \hat{D}(\mu_{k+1}, \Sigma_{k+1}, \beta_k) \right);$

5: **end for**

6: $\hat{\theta} \leftarrow (\mu_{K+1}, \Sigma_{K+1}, \beta_{K+1})$

In this algorithm, α denotes the step-size, which is selected according to the Armijo-Goldstein rule, and S_0 denotes a neighborhood of the true parameter value θ^* . The following Proposition 1 states the convergence of Algorithm 1.

Proposition 1. *Assume the cost function (31) has an isolated stationary point $\theta = \theta^*$ in some neighborhood $S_0 \subset \Theta$, and that the estimates $(\theta_k)_{k \geq 0}$ remain within S_0 . Then, for the sequence $(\theta_k)_{k \geq 0}$ generated by Algorithm 1,*

$$\lim_{k \rightarrow \infty} \theta_k = \theta^*$$

Appendix A sketches a proof of this convergence. For the case $\theta = (\Sigma)$ or $\theta = (\mu, \Sigma)$, near the true value θ^* , the Hessian of the function $\hat{D}(\theta)$ is approximated by the Fisher information metric. Therefore, one should expect the θ_k converge to θ^* with a superlinear rate of convergence, just like the Newton method dose. Precisely, if $\theta = (\Sigma)$ or $\theta = (\mu, \Sigma)$, with a fixed shape parameter β^* , then, under the assumptions of Proposition 1, one should expect Algorithm 1 to generate a sequence $(\theta_k)_{k \geq 0}$ converging superlinearly to θ^* . This is essentially due to Theorem 6.3.2 in [33], and will be observed experimentally in Section 5 (see Figure 1).

4.2. Stochastic gradient

The ISG method is an online quasi-Newton method. For each update, only one sample or one mini-batch is used. Here, the cost function remains the same as in equation (3). For the current estimate $\theta_k = (\mu_k, \Sigma_k, \beta_k)$ the stochastic information gradients are

$$\nabla_{\mu} \ell(\theta; x_n) = I_{\mu}^{-1} \Sigma G_{\mu}(x_n; \theta) \quad (33a)$$

$$\begin{aligned} \nabla_{\Sigma} \ell(\theta; x_n) = & J_1^{-1} [G_{\Sigma}(x_n; \theta)]^{\perp} \\ & + J_2^{-1} [G_{\Sigma}(x_n; \theta)]^{\parallel} \end{aligned} \quad (33b)$$

$$\nabla_{\beta} \ell(\theta; x_n) = I_{\beta}^{-1} G_{\beta}(x_n; \theta) \quad (33c)$$

Accordingly, the expected direction of descent is $\mathbb{E}_{\theta^*}[\nabla_{\theta}\ell(\theta; x)]$, which is equal to 0 at the global minimum θ^* . As in the classic stochastic gradient descent method, the step-size $\alpha_n = \frac{a}{n}$ is strictly positive, decreasing, and verifies the usual conditions

$$\sum \alpha_n = \infty \quad \sum \alpha_n^2 < \infty$$

Using the expressions (33a), (33b), (33c), the ISG algorithm can now be stated as follows.

Algorithm 2 ISG algorithm

Input: A dataset $\mathcal{X} = (x_1, \dots, x_N)$, an initialization $\theta_0 \in S_0 \subset \Theta$, the coefficient $a > 0$;

Output: The estimate $\hat{\theta}$;

```

1: for  $n = 0, 1, 2, \dots, N$  do
2:    $\alpha_{n+1} \leftarrow \frac{a}{n+1}$ ;
3:    $\mu_{n+1} \leftarrow \mu_n + \alpha_{n+1} \nabla_{\mu} \ell(\mu_n, \Sigma_n, \beta_n, x_n)$ ;
4:    $\Sigma_{n+1} \leftarrow \Sigma_n \exp(\Sigma_n^{-1} \alpha_{n+1} \nabla_{\Sigma} \ell(\mu_{n+1}, \Sigma_n, \beta_n, x_n))$ ;
5:    $\beta_{n+1} \leftarrow \beta_n \exp(\beta_n^{-1} \alpha_{n+1} \nabla_{\beta} \ell(\mu_{n+1}, \Sigma_{n+1}, \beta_n, x_n))$ ;
6: end for
7:  $\hat{\theta} \leftarrow (\mu_{N+1}, \Sigma_{N+1}, \beta_{N+1})$ ;

```

Remark that, the descending direction is $-\nabla_{\theta}\ell(\theta; x)$, and the double negative sign is simplified as positive in the algorithm. The compact and convex set S_0 is a neighborhood of θ^* , in which the cost function $D(\theta)$ has an isolated stationary point $\theta = \theta^*$. The following proposition 2 states the convergence of Algorithm 2.

Proposition 2. *Assume the function $D(\theta)$ has an isolated stationary point at $\theta = \theta^*$ in S_0 , and that the estimates $(\theta_n)_{n \geq 0}$ remain within S_0 . Then,*

$\lim \theta_n = \theta^*$ *almost surely*.

The proof of this convergence is discussed in Appendix B. Note that S_0 admits a system of normal coordinates $(\theta^i; i = 1, \dots, d)$ with origin at θ^* , where d is the dimension of the parameter space Θ , $d = \frac{m(m+1)}{2} + m + 1$. Since $D(\theta)$ has an isolated stationary point at $\theta = \theta^*$, the Hessian at point $\theta = \theta^*$ can be expressed in normal coordinates

$$\mathcal{H}_{ij} = \left. \frac{\partial^2 D}{\partial \theta^i \partial \theta^j} \right|_{\theta^i=0} \quad (34)$$

The matrix $\mathcal{H} = (\mathcal{H}_{ij})$ is positive definite [33]. With these notations, the rate of convergence is given by the following proposition.

Proposition 3. *Under the assumptions of Proposition 2, if $a > \frac{1}{2\lambda}$, where $\lambda > 0$ is the smallest eigenvalue of \mathcal{H} ,*

$$\mathbb{E}[\mathrm{d}^2(\theta^*, \theta_n)] = \mathcal{O}(n^{-1}) \quad (35)$$

Here, $\mathrm{d}(\cdot, \cdot)$ stands for the product distance in (28), and the "big O" notation means that there exist $K > 0$ and $n_0 > 0$ such that

$$\forall n \geq n_0 \quad \mathbb{E}[\mathrm{d}^2(\theta^*, \theta_n)] \leq \frac{K}{n}$$

In terms of the normal coordinates (θ^i) , let the direction of descent $\nabla_{\theta} \ell(\theta^*; x)$ at the point $\theta = \theta^*$ have components $(u^i(\theta^*))$. Let $\mathcal{G}^* = (\mathcal{G}_{ij}^*)$, be the matrix

$$\mathcal{G}_{ij}^* = \mathbb{E}_{\theta^*} [u^i(\theta^*) u^j(\theta^*)] \quad (36)$$

Then, the following proposition gives the asymptotic normality of the ISG algorithm.

Proposition 4 (asymptotic normality). *Under the assumptions of Propositions 2 and 3, the distribution of the re-scaled coordinates $(n^{\frac{1}{2}}\theta^i)_{i \in \{1, \dots, d\}}$ converges to a centred d -variate normal distribution, where d is the dimension of Θ , with covariance matrix \mathcal{G} given by the following Lyabunov equation*

$$A\mathcal{G} + \mathcal{G}A = -a^2\mathcal{G}^* \quad (37)$$

Here, $A = (A_{ij})$ with $A_{ij} = \frac{1}{2}\delta_{ij} - a\mathcal{H}_{ij}$ (δ denotes Kronecker's delta).

The proofs of Propositions 4 and 3 are discussed in Appendix C. For the case $\theta = (\Sigma)$ or $\theta = (\mu, \Sigma)$, the product metric (7) coincides with the information metric of the ECD model. Then, the assumptions of Proposition 5 in [28] are satisfied, and the following corollary may be obtained.

Corollary 1. *For the ECD model, parameterised by $\theta = (\Sigma)$ or $\theta = (\mu, \Sigma)$, with a fixed β^* , the product metric (7) coincides with the information metric.*

1. *the rate in equation (35) holds, whenever $a > 1/2$.*
2. *if $a = 1$ the distribution of the re-scaled coordinates $(n^{1/2}\theta^i)$ converges to a centred d -variate normal distribution, with covariance matrix equal to the identity $\mathcal{G}^* = I_d$, and the recursive estimates θ_n are asymptotically efficient.*

Note that, Item 2) of Corollary 1 implies that the distribution of $nd^2(\theta^*, \theta_n)$ converges to a χ^2 -distribution with d degrees of freedom.

$$nd^2(\theta^*, \theta_n) \Rightarrow \chi^2 \left(\frac{m(m+1)}{2} \right) \text{ for } \theta = (\Sigma) \quad (38a)$$

$$nd^2(\theta^*, \theta_n) \Rightarrow \chi^2 \left(\frac{m(m+1)}{2} + m \right) \text{ for } \theta = (\mu, \Sigma) \quad (38b)$$

This provides a practical means of confirming the asymptotic normality of the estimators θ_n . The function $d^2(\cdot, \cdot)$ denotes the square information distance, here the same as (28).

4.3. Global convergence analysis

This section studies the global convergence of the IDG and ISG algorithms, for two specific families of distributions, MGGD and Student-T. The main results are stated in the following two tables. For the cases indicated

Table 1: Convergence analysis: MGGD

	MGGD
$\theta = (\Sigma)$	Globally for $\beta > 0$
$\theta = (\mu, \Sigma)$	Globally for $\beta > 0$

Table 2: Convergence analysis: Student-T

	Student
$\theta = (\Sigma)$	Globally for $\beta > -m$
$\theta = (\mu, \Sigma)$	Globally for $\beta > 0$

in Tables 1 and 2, the cost function $D(\theta)$ (or $\hat{D}(\theta)$) has a unique stationary point, at θ^* , which is the global minimizer. This will be obtained from the following development.

First, for the case of $\theta = (\Sigma)$ with known μ^* and β^* , let

$$f(\delta_x, \beta) = \frac{1}{g(\delta_x, \beta)} \quad (39)$$

then, for the MGGD model

$$f(\delta_x, \beta) = \exp\left(\frac{1}{2}\delta_x^\beta\right) \quad \beta > 0 \quad (40a)$$

and for the Student-T model,

$$f(\delta_x, \beta) = \left(1 + \frac{\delta_x}{\beta}\right)^{\frac{\beta+m}{2}} \quad \beta > -m \quad (40b)$$

The following proposition introduces a sufficient condition for the KL divergence $D(\Sigma)$ and its empirical approximation $\hat{D}(\Sigma)$ to be geodesically strictly convex.

Proposition 5. *assume that the function $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ in (39) verifies the following condition : for any $\varphi : \mathbb{R} \rightarrow \mathbb{R}_+$*

$$\varphi \text{ strictly log-convex} \Rightarrow f \circ \varphi \text{ strictly log-convex} \quad (41)$$

Then, the KL divergence $D(\Sigma)$ (and its approximation \hat{D}) is geodesically strictly convex.

In particular, the unique global minimum, and the unique stationary point, of $D(\Sigma)$ is at the true Σ^* . This proposition 5 directly yields the following corollary, for the specific MGGD model and Student-T model, by plugging (40a) and (40b) into (41).

Corollary 2. *the KL divergence $D(\Sigma)$ and its empirical approximation $\hat{D}(\theta)$ are geodesically strictly convex, with unique global minimum (and unique stationary point), in both of the following cases.*

1. \mathcal{X} is distributed according to an MGGD model, with scatter matrix Σ^* and with shape parameter $\beta > 0$.

2. \mathcal{X} is distributed according to a Student-T model, with scatter matrix Σ^* and degree of freedom $\beta > -m$.

Thus, when Σ is unknown and β satisfies the conditions of Corollary 2, this corollary implies the global convergence of Algorithms 1 and 2. Precisely, these algorithms will always converge to the true value θ^* of the parameter θ .

For the more complicated situation $\theta = (\mu, \Sigma)$, global convergence does not always hold. The cost function $D(\theta)$ is not geodesically convex, but may be reformulated, using a new matrix argument [15]

$$S = \begin{bmatrix} \Sigma + \mu\mu^\dagger & \mu \\ \mu^\dagger & 1 \end{bmatrix} \quad (42)$$

If the new random vector y is given by

$$y = (x^\dagger, 1)^\dagger \quad (43)$$

then the cost function can be reformulated as

$$\tilde{D}(\theta) = -\frac{1}{2} \log \det(S) - \log \tilde{f}(\delta_y) \quad (44)$$

where

$$\delta_y = y^\dagger S^{-1} y = (x - \mu)^\dagger \Sigma^{-1} (x - \mu) + 1 \quad (45)$$

and

$$\tilde{f}(\delta_y) = \exp \left[\frac{1}{2} (\delta_y - 1)^\beta \right] \text{ for MGGD} \quad (46a)$$

Then for Student-T is

$$\tilde{f}(\delta_y) = \left(1 - \frac{1}{\beta} + \frac{\delta_y}{\beta} \right)^{\frac{\beta+m}{2}} \text{ for Student-T} \quad (46b)$$

In [15], the minimization of $\tilde{D}(\theta)$ was proven to be equivalent to the minimization of $D(\theta)$. Replacing the new function \tilde{f} into (41), the following corollary is obtained.

Corollary 3. *the KL divergence $D(\mu, \Sigma)$ (and $\hat{D}(\mu, \Sigma)$) has a unique global minimum (and unique stationary point) at (μ^*, Σ^*) , in both of the following cases.*

1. *\mathcal{X} is distributed according to an MGGD model, with expectation and scatter matrix (μ^*, Σ^*) and with fixed shape parameter $\beta > 0$.*
2. *\mathcal{X} is distributed according to a Student-T model, with expectation and scatter matrix (μ^*, Σ^*) and with the fixed degree of freedom $\beta > 0$.*

For these two cases, global convergence is then guaranteed.

Finally, for the most complicated case, $\theta = (\mu, \Sigma, \beta)$, the cost function is always non-convex. Moreover, we have verified experimentally that it has multiple stationary points in $\Theta = \mathbb{R}^m \times \mathcal{P}_m \times \mathbb{R}_+$. Therefore, the correct estimation can only be guaranteed when the initial value θ_0 is close enough to the global minimum θ^* .

5. Computer experiments

This section presents a set of computer experiments, which confirm the theoretical results of Section 4, and provide a detailed comparison of the ISG and IDG estimation methods, with the already existing MM and FP. For every experiment, 1000 Monte Carlo trials were carried out. For each trial, the dataset $\mathcal{X} = \{x_1, \dots, x_N\}$ is independent and identically distributed,

according to true parameters $(\mu^*, \Sigma^*, \beta^*)$. The dimension m of x_n is taken equal to 10. The true μ^* is randomly chosen from a multivariate normal distribution. The scatter Σ^* is defined as $\Sigma(i, j) = \rho^{|i-j|}$ for $i, j \in \{1, m\}$, and $\rho \sim \mathcal{U}(0.2, 0.8)$. The shape parameter β^* is uniformly selected from the intervals $[0.2, 5]$ for MGGD and for Student-T.

The first experiment confirms the super-linear rate of convergence of IDG, for a dataset, distributed according to the MGGD model, which contains $N = 10^4$ samples. The initial value θ_0 is defined as the MM estimate, using 10% of the entire dataset. Figure 1(a) presents the case of $\theta = (\Sigma)$ with known (μ^*, β^*) . The IDG method converges after only two iterations, and if the same accuracy needs to be achieved, the deterministic gradient method (not using the information gradient) requires at least 88 iterations. For the case of $\theta = (\mu, \Sigma)$ with known (β^*) , things are similar. Figure 1(b) shows that IDG, after two iterations, achieves the same accuracy as the traditional gradient method, after 200 iterations.

The second experiment confirms the convergence rate of ISG. In this experiment, both MGGD and Student-T datasets are used. The initialization θ_0 is randomly chosen. Figures 2(a), 2(b), and 2(c) confirm the rate of convergence stated in (35), in the neighborhood of $\theta^* = (\mu^*, \Sigma^*, \beta^*)$. In these log-log plots, the x-axis and y-axis represent the number of iterations and $\mathbb{E}[d^2(\theta^*, \hat{\theta}_n)]$, respectively, and \mathbb{E} denotes the Monte Carlo approximation of the expectation, obtained by averaging over the 1000 trials. The slope of each curve approaches -1 , while θ_n approaches the true value θ^* . Note that, for the cases of $\theta = (\Sigma)$ and $\theta = (\mu, \Sigma)$, the initialization θ_0 can be chosen far away from θ_0 (e.g. $d^2(\theta^*, \theta_0) > 10$). However, when $\theta =$

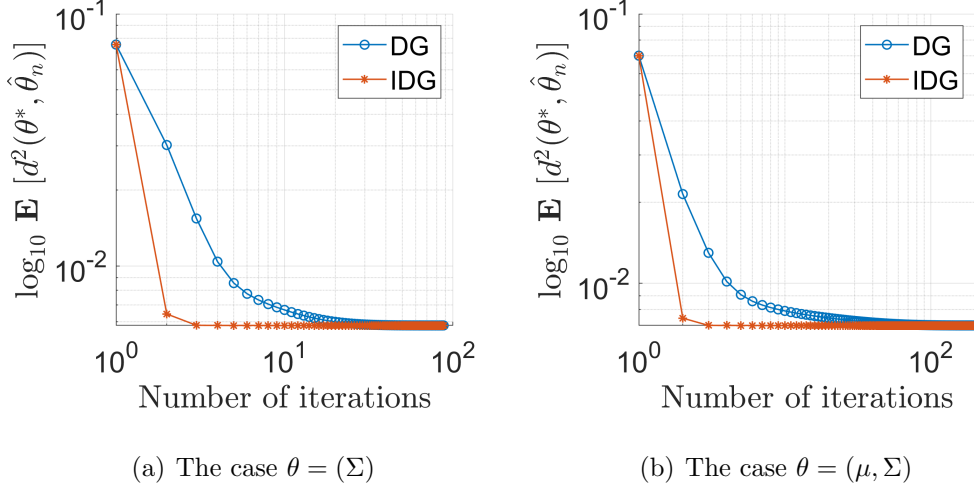
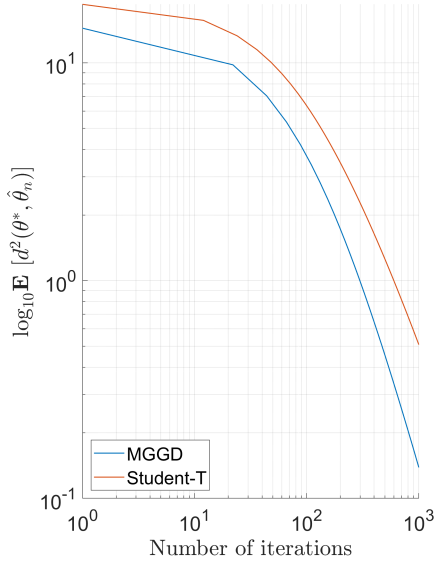


Figure 1: The superlinear convergence rate for IDG

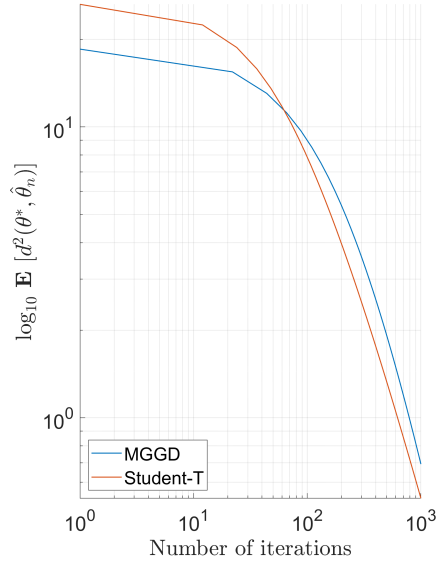
(μ, Σ, β) , the initialization should be in a neighborhood of θ^* which satisfies the conditions in Proposition 3. For the results obtained in Figures 2(a) and 2(b) (that is to say, when β^* is fixed), the step-size coefficient a always equals 1, which satisfies the condition in 1. For the case of unknown β , the step-size coefficient a is taken much larger, in order to meet the conditions of proposition 3. In fact, here, $a = 100$.

For the case of $\theta = (\Sigma)$ and $\theta = (\mu, \Sigma)$, Figures 3(a) and 3(b) confirm the chi-squared limit distribution in corollary (1). The samples x_n being matrices of size $m \times m$ with $m = 10$, the dashed blue curve is the probability density of a chi-squared distribution with 55 and 65 degrees of freedom, for Figures 3(a) and 3(b) respectively. The solid lines are the smoothed histograms of $Nd^2(\theta^*, \hat{\theta}_N)$ where $N = 10^5$. These "estimated p.d.f." coincide very closely with the theoretical chi-squared probability density.

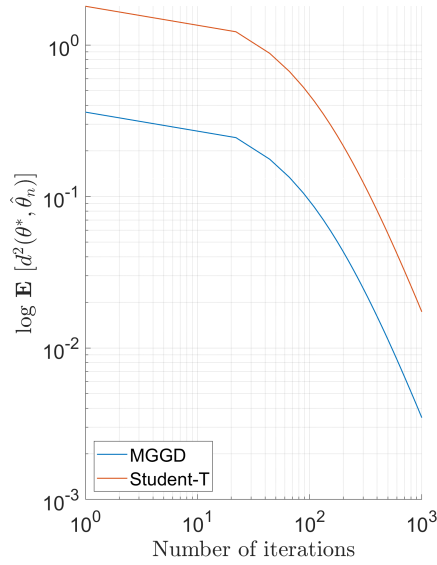
In the third experiment, we compare the efficiency of the IDG and ISG



(a) The case $\theta = (\Sigma)$



(b) The case $\theta = (\mu, \Sigma)$



(c) The case $\theta = (\mu, \Sigma, \beta)$

Figure 2: Linearly convergence rate for ISG

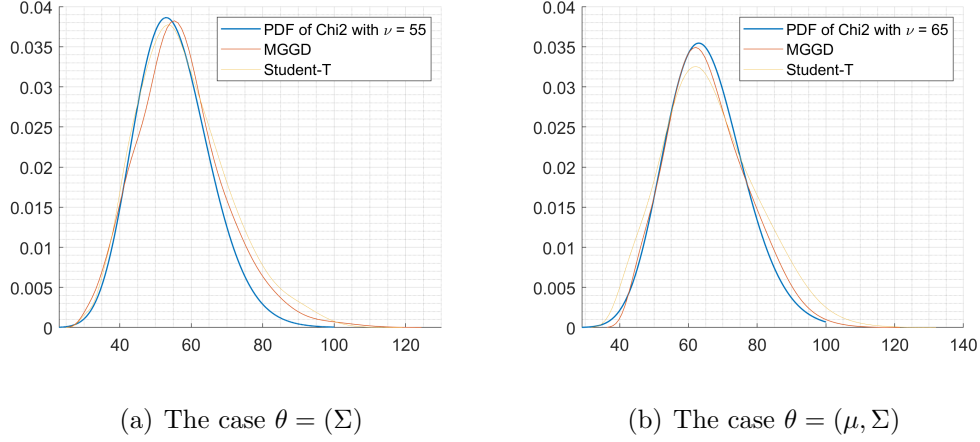
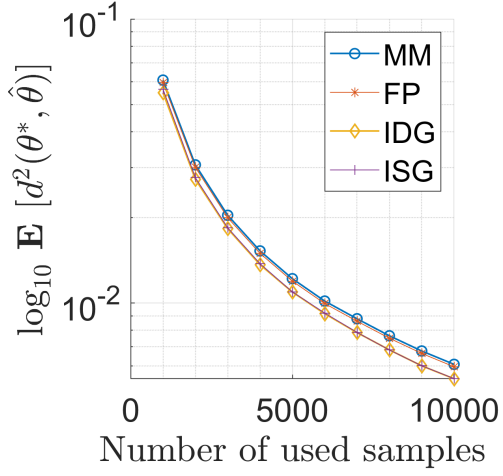
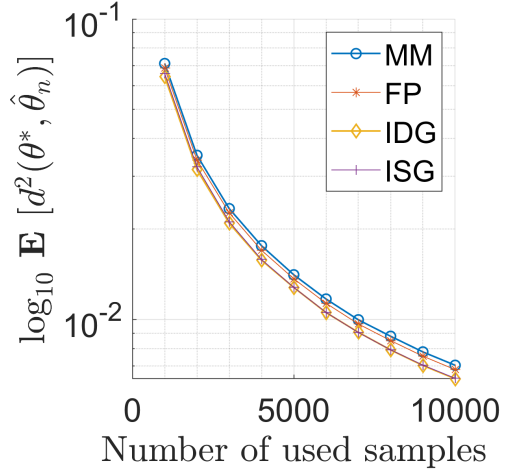


Figure 3: Validation by fitting a Chi-2

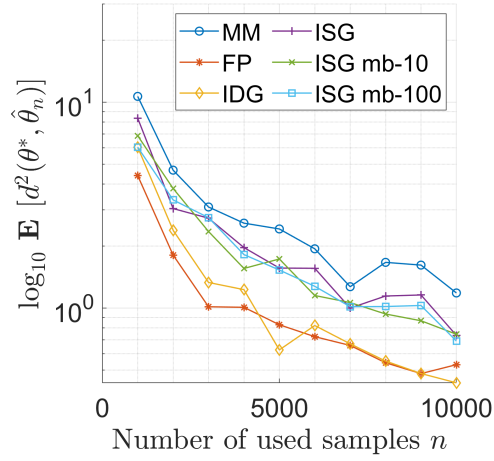
methods with other common estimation methods, MM and FP. In each trial, the dataset is generated from an MGGD model, and contains $N = 10^4$ datapoints. For MGGD, the MM was given in [25], and the FP method in [24]. In Figures 4(a), 4(b) and 4(c), the x-axis denotes the size of the dataset, and the y-axis denotes the expectation of the square distance between θ^* and the estimated $\hat{\theta}$. This expectation is approximated by the average of 10^3 Monte Carlo trials. For the cases $\theta = (\Sigma)$ and $\theta = (\mu, \Sigma)$, the IDG and ISG algorithms show a better accuracy. When $\theta = (\mu, \Sigma, \beta)$, the accuracy of the MLE method is still significantly better than MM, and the accuracies of IDG and FP coincide. However, the accuracy of ISG is not as good as as FP or IDG. This phenomenon may be explained theoretically. Indeed, when $\theta = (\mu, \Sigma, \beta)$, the product metric does not coincide with the information metric of the ECD model, and this leads to a less efficient estimation. The fluctuations of the curves in Figure 4(c) are quite significant. This means the variance of the final estimate $\hat{\theta}$ is significant.



(a) The case $\theta = (\Sigma)$



(b) The case $\theta = (\mu, \Sigma)$

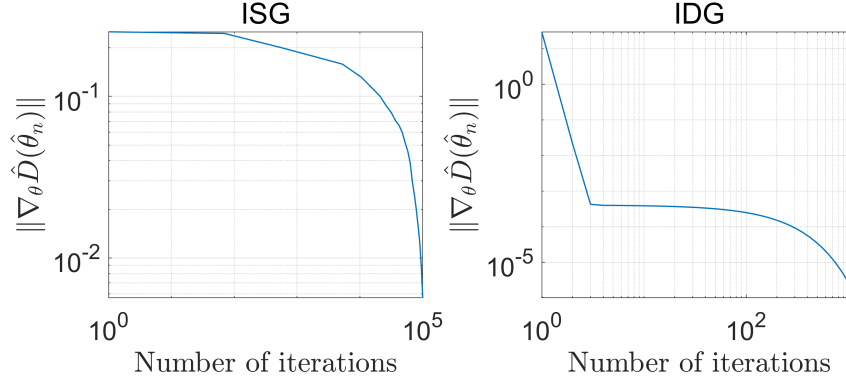


(c) The case $\theta = (\mu, \Sigma, \beta)$

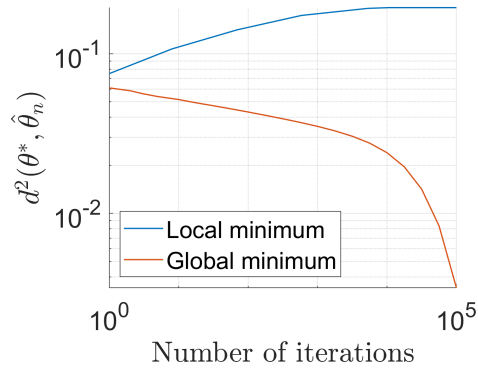
Figure 4: Efficiency comparison

Two additional experiments were done, to explain these fluctuations. The first additional experiment shows that both IDG and ISG eventually converge to a stationary point (but not necessarily a global minimum). The variations of the norms of the gradients appear in Figure 5(a). As the number of iterations increases, the norm of the gradient approaches 0, for both IDG and ISG. The second additional experiment proved the existence of stationary points other than the true value θ^* . For the same dataset, two different initial values θ_0 were used for the ISG method. In Figure 5(b), the initial value θ_0 of the red curve is close to the global minimum θ^* , and its θ_n finally converge to θ^* . The blue curve has θ_0 farther away, and its $d^2(\theta^*, \hat{\theta}_n)$ converges to a non-zero constant. In conclusion, for $\theta = (\mu, \Sigma, \beta)$, the convergence to global minimum θ^* can only be guaranteed locally. If the initial value θ_0 is chosen in a neighborhood S_0 , then FP and IDG can converge to the true point by virtue of their stability, where S_0 should always satisfy the conditions in proposition 1 and 2. Due to its stochastic nature, ISG may jump out of the neighborhood S_0 during the first few iterations. This leads to convergence to local minimum, different from θ^* . Then, the final averaged accuracy of ISG is not as good as the other two MLE methods, and the variance of the ISG estimator is relatively important. As a possible remedy to this problem, the mini-batch ISG was also tested, and compared with other methods, in the Figure 4(c). Two sizes of the mini-batch, 10 and 100, were considered. However, the experimental results show that the mini-batch has no significant effect on the accuracy of ISG.

As for computational time, information gradient methods have a significant advantage. The computational time of the ISG algorithm is similar to



(a) Gradient $\rightarrow 0$.

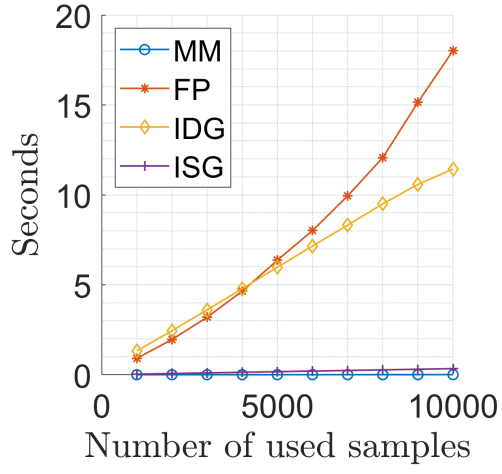


(b) Other stationary point

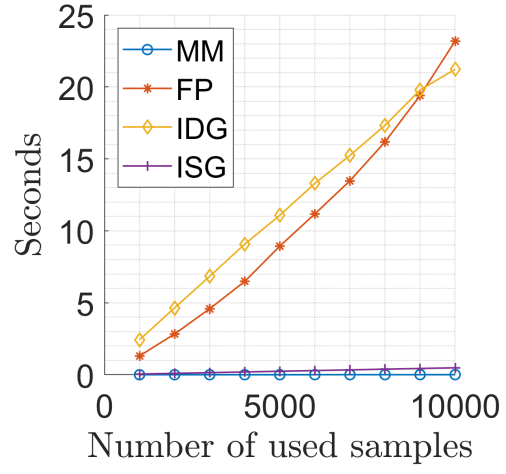
Figure 5: Additional experiments

Table 3: Percentage of 'correct' estimates

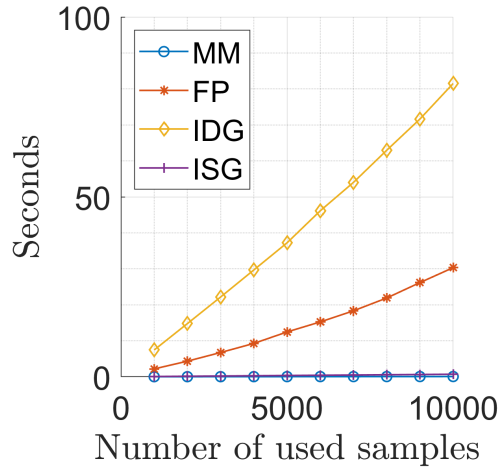
	correct estimates $d^2(\theta^*, \theta_n) \searrow 0$ and $\nabla_{\theta} D(\theta_n) \searrow 0$	incorrect estimates $d^2(\theta^*, \theta_n) \rightarrow c \gg 0$ and $\nabla_{\theta} D(\theta_n) \searrow 0$
$\theta = (\mu, \Sigma, \beta)$	73%	27%



(a) The case $\theta = (\Sigma)$



(b) The case $\theta = (\mu, \Sigma)$



(c) The case $\theta = (\mu, \Sigma, \beta)$

Figure 6: Time consumption

that of MM, and is significantly less than that of FP. Meanwhile, its accuracy is significantly better than that of MM. In most experiments, the accuracy of ISG is similar to, or even better than, FP. Although the computational time of IDG is greater than that of ISG, it is comparable to that of FP, while, in most cases, IDG can achieve the best accuracy, among the four estimation methods considered.

6. Application with real dataset

In addition to experimental simulations, we also applied our methods to real datasets.

6.1. Color transformation

The first application is to color transformation for image editing with MGGD models, which was investigated in [37]. Its goal is to replace the color distribution of the input image by that one of a target image. The main idea is to fit the input and the target distributions, with two different MGGD models. Then, the transformation between these two MGGDs is implemented by a linear Monge-Kantorovich transformation for Σ , and a stochastic transformation for β . Specifically, this conversion can be three-dimensional (3D), for RGB images, or five-dimensional (5D), when spatial gradient-field information is included.

Starting with the 3D rgb case, Figure 7 presents the transformed images and some of their details. The detail (a1) clearly shows that the cloud 'drawn' by MM appears too green. Similarly, FP also presents a green appearance, in detail (a2). On the contrary, the two gradient methods, i.e. IDG and ISG methods, show pure white cloud color in (a3) and (a4). Note also the

difference in the amount of blue in the shadows on the grass. Too much blue is mixed with the shadow, in MM’s output detail (b1). In details (b2),(b3),(b4), the results of MLE methods lead to a more natural appearance.

From the point of view of the present work, the most interesting aspect of this application is in term of computational time. The recursive (online) ISG method takes about 10 seconds for two images (input and output). In contrast, FP and IDG each require more than two hours. In other words, ISG has a decisive advantage, in terms of time consumption.

Then, gradient-field information was included, so the transformation came to involve 5D, which consist in three color components (of CIELAB) and two components of the image spatial gradient field (dx and dy). For this application, the shape parameter of the MGGD model was supposed to fixed. Figure 8 presents the four different implementations. It can be observed that the output of the three MLE methods is significantly better than that of MM. In the transformed result of MM, the hue is darker and greener. MLE results are better, since the frost on the grass is whiter and appears more natural, and the forest on the mountain in the image also appears darker. The two images in Figure 8 have more than 1.2×10^6 pixels (i.e. 1.2×10^6 samples). The FP and IDG need more than 4 hours to run, on the these two images. The ISG method needs only 21 seconds.

We also considered an application to full HD images. In this case, as demonstrated in Figure 9, the advantages of the ISG algorithm were significant. The result of MM failed to achieve the color of the autumn leaves in the target image, showing cyan instead of yellow. Since the input image and the target image have more than 4×10^6 pixels (that is 4×10^6 samples), it



Figure 7: 3D transformation



Figure 8: 5D transformation of images that have moderate size

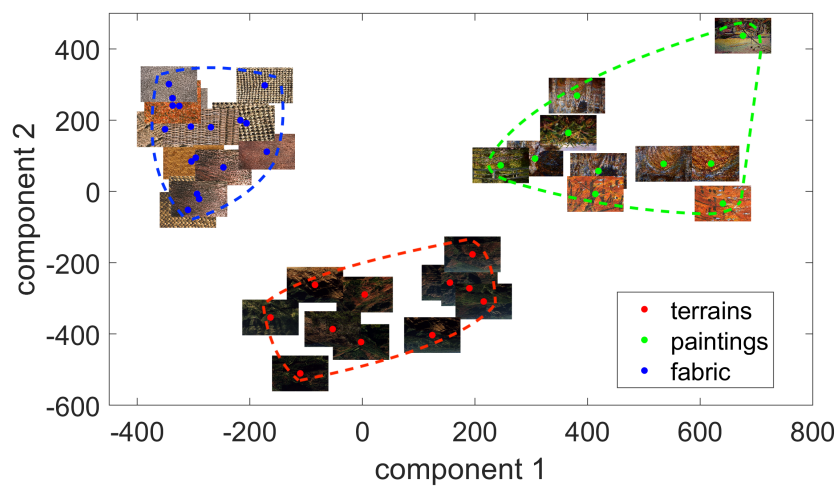
was not feasible to run FP and IDG, with the entire dataset. Rather, the estimation was done on subsets of the complete dataset. These two subsets have 4×10^5 samples, that are randomly taken from the original images. In the autumn leaves obtained using FP and IDG, the yellow color has obviously been smeared. ISG is more natural, in which the yellow color is more uniform, and it is closer to the style of the target image.

6.2. Classification

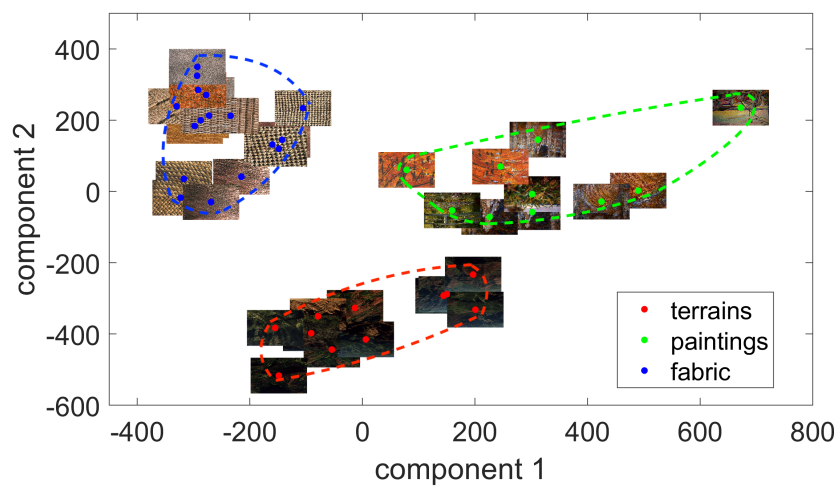
MGGD are also used for texture modeling [38, 25]. Without going into the details of presently existing classification methods, we attempted to use an MGGD representation, in order to distinguish between different groups of textures. Three groups of textures are selected from the VisTex database [39], 11 paintings, 18 fabrics, and 11 terrains. Each texture is considered as an RGB 3-dimensional image, modeled by an MGGD, whose parameters $\theta = (\mu, \Sigma, \beta)$ are estimated by two MLE methods, i.e. FP and ISG. Then, the scatter matrices Σ are normalized by their trace, i.e. $M = \frac{p}{\text{tr}(\Sigma)}\Sigma$ (in order to



Figure 9: 5D transformation of images that have large size



(a) FP



(b) ISG

Figure 10: Texture Classification using MGGD

avoid the elements of Σ being too small). Afterwards, for each texture, a new vector is constituted by the eigenvalues of M , μ and β . This 7-dimensional vector is projected onto a 2D plane, via a PCA operation. A visual (2D) representation is given in Figure 10. These two figures are obtained using FP and ISG, respectively. Each texture contains $512 \times 512 = 262144$ pixels, FP expended 210 seconds for each image. In stark contrast, ISG expended 4.6 seconds in average. And for both these methods, the boundary between these 3 clouds of points is quite sharp, and the distinction is quite clear. We have reason to believe that, in this scenario, ISG has achieved the same performance as FP and simultaneously it used less time.

Acknowledgement

At the end, we thank the support from the ANR MAR- GARITA under Grant ANR-17-ASTR-0015 for our works.

Appendix A. Proof of proposition 1

Let $(\theta_k)_{k \geq 0}$ be an infinite sequence generated by Algorithm 1. Recall the retraction R_θ defined in (24). Consider the sequence of tangent vectors $(\eta_k)_{k \geq 0}$ where η_k belongs to $T_{\theta_k} \Theta$, and $\eta_0 = -\nabla_\mu \hat{D}(\mu_0, \Sigma_0, \beta_0)$, $\eta_1 = -\nabla_\Sigma \hat{D}(\mu_1, \Sigma_0, \beta_0)$, $\eta_2 = -\nabla_\beta \hat{D}(\mu_1, \Sigma_1, \beta_0)$, and so on.

Then, the sequence (θ_k) is given as in Algorithm 1 of [33], $\theta_{k+1} = R_{\theta_k}(t_k \eta_k)$ with step-size t_k chosen according to Armijo-Goldstein rule (note that $t_0 = \alpha_\mu$, $t_1 = \alpha_\Sigma$, and $t_2 = \alpha_\beta$, etc.).

The sequence (θ_k) remains within the neighborhood S_0 of θ^* . Without loss of generality, assume this neighborhood is compact. According to Corollary

4.3.2 in [33], if the sequence $(\eta_k)_{k \geq 0}$ is gradient-related,

$$\lim_{k \rightarrow \infty} \|\text{grad} \hat{D}(\theta_k)\| = 0 \quad (\text{A.1})$$

Then, since θ^* is the only stationary point of the cost function (31) in S_0 , it follows that $\lim_{k \rightarrow \infty} \theta_k = \theta^*$, as required. To show that the sequence $(\eta_k)_{k \geq 0}$ is gradient-related, note that

$$\begin{aligned} \langle \eta_0, \nabla_{\theta} \hat{D}(\theta_0) \rangle &= -\|\nabla_{\mu} \hat{D}(\mu_0, \Sigma_0, \beta_0)\|^2 \\ \langle \eta_1, \nabla_{\theta} \hat{D}(\theta_1) \rangle &= -\|\nabla_{\Sigma} \hat{D}(\mu_1, \Sigma_0, \beta_0)\|^2 \\ \langle \eta_2, \nabla_{\theta} \hat{D}(\theta_2) \rangle &= -\|\nabla_{\beta} \hat{D}(\mu_1, \Sigma_1, \beta_0)\|^2 \end{aligned}$$

and so on, for $k \geq 3$. In other words, the scalar product between η_k and $\nabla_{\theta} \hat{D}(\theta_k)$ is always strictly negative. Therefore, the sequence $(\eta_k)_{k \geq 0}$ is gradient-related.

Appendix B. Proof of proposition 2

The proof is a direct application of Remark 2, concerning Proposition 1, in [28]. According to this remark, if $u(\theta_n, x)$ denotes the direction of descent, and if

$$\mathbb{E} \langle u(\theta_n, x), \nabla_{\theta} D(\theta_n) \rangle < 0, \text{ for } n > 0 \quad (\text{B.1})$$

then $\lim \theta_n = \theta^*$ almost surely. Here (compare to the proof of Proposition 1), the direction of descent is given by $u(\theta_0, x) = \nabla_{\mu} \ell(\mu_0, \Sigma_0, \beta_0)$, $u(\theta_1, x) = \nabla_{\Sigma} \ell(\mu_1, \Sigma_0, \beta_0)$, $u(\theta_2, x) = \nabla_{\beta} \ell(\mu_1, \Sigma_1, \beta_0)$ and so on. Therefore, the expectation in (B.1) is equal to

$$\mathbb{E} \langle u(\theta_0, x), \nabla_{\theta} D(\theta_0) \rangle = -\|\nabla_{\mu} D(\mu_0, \Sigma_0, \beta_0)\|^2$$

$$\mathbb{E} \langle u(\theta_1, x), \nabla_\theta D(\theta_1) \rangle = -\|\nabla_\Sigma D(\mu_1, \Sigma_0, \beta_0)\|^2$$

$$\mathbb{E} \langle u(\theta_2, x), \nabla_\theta D(\theta_2) \rangle = -\|\nabla_\beta D(\mu_1, \Sigma_1, \beta_0)\|^2$$

and so on, for $k \geq 3$. This shows that (B.1) is verified.

Appendix C. Proof of propositions 3 and 4

As for Proposition 2, this is an application of Remark 2 in [28]. According to this remark, in order to obtain the mean-square rate and the asymptotic normality, it is enough to show the mean vector field $X(\theta) = \mathbb{E}_{\theta^*}[u(\theta, x)]$ has an attractive stationary point at $\theta = \theta^*$. Since $u(\theta, x) = \nabla_\theta \ell(\theta, x)$

$$\mathbb{E}_{\theta^*}[u(\theta; x)] = \begin{bmatrix} \nabla_\mu D(\theta) \\ \nabla_\Sigma D(\theta) \\ \nabla_\beta D(\theta) \end{bmatrix} \quad (\text{C.1})$$

The covariant derivative of this vector field at the point $\theta = \theta^*$ is equal to the Hessian $\mathcal{H}(\theta^*)$, which is positive definite. Therefore, the results of Propositions 3 and 4 follow by Remark 2 in [28].

Appendix D. Proof of proposition 5

For the case of $\theta = (\Sigma)$, the geodesic convexity of the cost function $D(\theta)$ (or of $\hat{D}(\theta)$) follows by proving $-\ell(\theta; x)$ is geodesically strictly convex in $\theta = (\Sigma)$ for any x .

To do this, for any fixed x , denote $g(\theta) = -\ell(\theta; x)$. Recall that, geodesic curves on \mathcal{P}_m are of the form [35]

$$\begin{aligned} \gamma : \mathbb{R} &\rightarrow \mathcal{P}_m \\ t &\mapsto A \exp(tr) A^\dagger \end{aligned} \quad (\text{D.1})$$

where \exp denotes the matrix exponential map, A is an invertible matrix, and r is a diagonal matrix, both of same size as Σ . Then, $g(\theta)$ is geodesically convex if and only if the composition $(g \circ \gamma)(t)$ is always a convex function with respect to t . Moreover, geodesic strict convexity is defined in exactly the same way. The composition $(g \circ \gamma)(t)$ can be expressed

$$(g \circ \gamma)(t) = \log \det(A) + \text{tr}(r) \frac{t}{2} + \log [(f \circ \varphi)(t)] \quad (\text{D.2})$$

where

$$\varphi(t) = \sum_{i=1}^p u_i^2 \exp(-r_i t) \quad (\text{D.3})$$

$u = A^{-1}x$ has components u_i , and r_i are the diagonal elements of r . The function $\varphi : \mathbb{R} \rightarrow \mathbb{R}_+$ is strictly log-convex, because it is the Laplace transform of a positive measure [40]

$$\varphi(t) = \int_0^\infty \exp(-tx) \mu(dx) \quad (\text{D.4})$$

where $\mu = \sum_{i=1}^m u_i^2 \delta_{r_i}$, and δ_{r_i} is the Dirac measure concentrated at r_i .

Assume that the function f verifies Condition (41). Then, since φ is strictly log-convex, $f \circ \varphi$ is strictly log-convex. Thus, the term $\log [(f \circ \varphi)(t)]$ of (D.2) is a strictly convex function of the real variable t . Since the term $\text{tr}(r) \frac{t}{2}$ of (D.2) amounts to an affine function of t , it is now clear that $(g \circ \gamma)(t)$ is a strictly convex function of the real variable t , for any geodesic curve $\gamma : \mathbb{R} \rightarrow \mathcal{P}_m$. Finally, since x was chosen arbitrarily, $-\ell(\theta; x)$ is geodesically strictly convex in $\theta = (\Sigma)$ for each x . Therefore, $D(\theta)$ and $\hat{D}(\theta)$ are both geodesically strictly convex.

Appendix E. Proof of corollary 2 and 3

For the case of $\theta = (\Sigma)$, note that $\varphi : \mathbb{R} \rightarrow \mathbb{R}_+$ is strictly log-convex if and only if $\varphi(t) = \exp(\psi(t))$ where $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is strictly convex.

1) plugging (40a) into (41),

$$\log(f \circ \varphi)(t) = \frac{1}{2} \exp(\beta(\psi(t))) \quad (\text{E.1})$$

Therefore, condition (41) is verified since $\beta > 0$.

2) plugging (40b) into (41),

$$\log(f \circ \varphi)(t) = \frac{\beta + m}{2} \log\left(1 + \frac{\exp(\psi(t))}{\beta}\right) \quad (\text{E.2})$$

Therefore, condition (41) is verified since $\beta + m > 0$.

For the case of $\theta = (\mu, \Sigma)$, as mentioned above, the function \tilde{f} is reformulated. Then, the same strategy is applied for this reformulated \tilde{f} .

1) For MGGD, recall the geodesic curve for reformulated matrix $S(t)$,

$$S(t) = B \exp(st) B^\dagger \quad (\text{E.3})$$

where \exp denotes the matrix exponential map, B is an invertible matrix, and s is a diagonal matrix, both of same size as S .

$$\delta_y(t) = y^\dagger S^{-1} y = \sum_{i=1}^{p+1} v_i^2 e^{-s_i t} \quad \text{with } v = B^{-1} y \quad (\text{E.4})$$

According to equation (45), we have $\delta_y > 1$. Therefore, $\exists w \in \mathbb{R}^{p+1}$ and $\exists q \in (0, +\infty)^{p+1}$ (e.g. $w = (u, 0)$ and $q = (r, 1)$) such that

$$\sum_{i=1}^{p+1} v_i^2 e^{-s_i t} = \sum_{i=1}^{p+1} w_i^2 e^{-q_i t} + 1 \quad (\text{E.5})$$

Plugging $\sum_{i=1}^{p+1} w_i^2 e^{-q_i t} + 1$ into the reformulated \tilde{f}

$$\tilde{f} \circ \delta_y(t) = \exp \left\{ \frac{1}{2} \left(\sum_{i=1}^{p+1} w_i^2 e^{-q_i t} \right)^\beta \right\} \quad (\text{E.6})$$

This function is proved to be log-convex in equation (D.3). Therefore, condition (41) is verified since $\beta > 0$ for MGGD model.

2) For Student-T, plugging (46b) into (41),

$$\log \left(\tilde{f} \circ \varphi \right) (t) = \frac{\beta + m}{2} \left[1 - \frac{1}{\beta} + \frac{1}{\beta} \exp(\psi(t)) \right] \quad (\text{E.7})$$

Therefore, condition (41) is verified since $\beta > 0$.

References

- [1] D. Kelker, Distribution theory of spherical distributions and a location-scale parameter generalization, *Sankhyā: The Indian Journal of Statistics, Series A* (1970) 419–430.
- [2] K. Fang, Y. Zhang, *Generalized multivariate analysis*, Science Press, 1990.
- [3] K. W. Fang, *Symmetric multivariate and related distributions*, CRC Press, 2018.
- [4] E. Gómez, M. Gomez-Vilegas, J. Marin, A multivariate generalization of the power exponential family of distributions, *Communications in Statistics-Theory and Methods* 27 (1998) 589–600.
- [5] E. G. Sánchez-Manzano, M. A. Gomez-Villegas, J.-M. Marín-Diazaraque, A matrix variate generalization of the power exponential

- family of distributions, *Communications in Statistics-Theory and Methods* 31 (2002) 2167–2182.
- [6] S. Kotz, S. Nadarajah, *Multivariate t-distributions and their applications*, Cambridge University Press, 2004.
 - [7] L. Boubchir, J. M. Fadili, Multivariate statistical modeling of images with the curvelet transform, in: *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications*, volume 2, IEEE, 2005, pp. 747–750.
 - [8] D. Cho, T. D. Bui, G. Chen, Image denoising based on wavelet shrinkage using neighbor and level dependency, *International journal of wavelets, multiresolution and information processing* 7 (2009) 299–311.
 - [9] G. Verdoolaege, S. De Backer, P. Scheunders, Multiscale colour texture retrieval using the geodesic distance between multivariate generalized gaussian models, in: *2008 15th IEEE International Conference on Image Processing*, IEEE, 2008, pp. 169–172.
 - [10] Y. Bazi, L. Bruzzone, F. Melgani, Image thresholding based on the em algorithm and the generalized gaussian distribution, *Pattern Recognition* 40 (2007) 619–634.
 - [11] J. Scharcanski, A wavelet-based approach for analyzing industrial stochastic textures with applications, *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37 (2006) 10–22.

- [12] P. Gupta, A. K. Moorthy, R. Soundararajan, A. C. Bovik, Generalized gaussian scale mixtures: A model for wavelet coefficients of natural images, *Signal Processing: Image Communication* 66 (2018) 87 – 94.
- [13] M. N. Desai, R. S. Mangoubi, Robust gaussian and non-gaussian matched subspace detection, *IEEE Transactions on Signal Processing* 51 (2003) 3115–3127.
- [14] T. Elguebaly, N. Bouguila, Bayesian learning of generalized gaussian mixture models on biomedical images, in: *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, Springer, 2010, pp. 207–218.
- [15] F. Laus, G. Steidl, Multivariate myriad filters based on parameter estimation of student-t distributions, *SIAM Journal on Imaging Sciences* 12 (2019) 1864–1904.
- [16] L. Bombrun, G. Vasile, M. Gay, F. Totir, Hierarchical segmentation of polarimetric sar images using heterogeneous clutter models, *IEEE Transactions on Geoscience and Remote Sensing* 49 (2011) 726–737.
- [17] J. I. Fernández-Michelli, M. Hurtado, J. A. Areta, C. H. Muravchik, Unsupervised polarimetric sar image classification using \mathcal{G}_p^0 mixture model, *IEEE Geoscience and Remote Sensing Letters* 14 (2017) 754–758.
- [18] Q. Chen, H. Yang, L. Li, X. Liu, A novel statistical texture feature for sar building damage assessment in different polarization modes, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020) 154–165.

- [19] D. E. Tyler, A distribution-free m-estimator of multivariate scatter, *The annals of Statistics* (1987) 234–251.
- [20] E. Ollila, D. E. Tyler, V. Koivunen, H. V. Poor, Complex elliptically symmetric distributions: Survey, new results and applications, *IEEE Transactions on Signal Processing* 60 (2012) 5597–5625.
- [21] S. Sra, R. Hosseini, Geometric optimisation on positive definite matrices for elliptically contoured distributions, in: *Advances in Neural Information Processing Systems*, 2013, pp. 2562–2570.
- [22] T. Zhang, A. Wiesel, M. S. Greco, Multivariate generalized gaussian distribution: Convexity and graphical models, *IEEE Transactions on Signal Processing* 61 (2013) 4141–4148.
- [23] S. Sra, R. Hosseini, Conic geometric optimization on the manifold of positive definite matrices, *SIAM Journal on Optimization* 25 (2015) 713–739.
- [24] F. Pascal, L. Bombrun, J.-Y. Tournier, Y. Berthoumieu, Parameter estimation for multivariate generalized gaussian distributions, *IEEE Transactions on Signal Processing* 61 (2013) 5960–5971.
- [25] G. Verdoolaege, P. Scheunders, Geodesics on the manifold of multivariate generalized gaussian distributions with an application to multi-component texture discrimination, *International Journal of Computer Vision* 95 (2011) 265.
- [26] S. Bonnabel, Stochastic gradient descent on riemannian manifolds, *IEEE Transactions on Automatic Control* 58 (2013) 2217–2229.

- [27] N. Tripuraneni, N. Flammarion, F. Bach, M. I. Jordan, Averaging stochastic gradient descent on riemannian manifolds, arXiv preprint arXiv:1802.09128 (2018).
- [28] J. Zhou, S. Said, Fast, asymptotically efficient, recursive estimation in a riemannian manifold, *Entropy* 21 (2019) 1021.
- [29] S.-i. Amari, Information geometry and its applications, volume 194, Springer, 2016.
- [30] S.-I. Amari, Natural gradient works efficiently in learning, *Neural computation* 10 (1998) 251–276.
- [31] G. Verdoolaege, P. Scheunders, On the geometry of multivariate generalized gaussian models, *Journal of mathematical imaging and vision* 43 (2012) 180–193.
- [32] J. C. Fuhrer, G. R. Moore, S. D. Schuh, Estimating the linear-quadratic inventory model maximum likelihood versus generalized method of moments, *Journal of Monetary Economics* 35 (1995) 115–157.
- [33] P.-A. Absil, R. Mahony, R. Sepulchre, Optimization algorithms on matrix manifolds, Princeton University Press, 2009.
- [34] M. Berkane, K. Oden, P. M. Bentler, Geodesic estimation in elliptical distributions, *Journal of Multivariate Analysis* 63 (1997) 35–46.
- [35] X. Pennec, P. Fillard, N. Ayache, A riemannian framework for tensor computing, *International Journal of computer vision* 66 (2006) 41–66.

- [36] C. Mostajeran, R. Sepulchre, Ordering positive definite matrices, *Information Geometry* 1 (2018) 287–313.
- [37] H. Hristova, O. Le Meur, R. Cozot, K. Bouatouch, Transformation of the multivariate generalized gaussian distribution for image editing, *IEEE transactions on visualization and computer graphics* 24 (2017) 2813–2826.
- [38] R. Kwitt, P. Meerwald, A. Uhl, G. Verdoolaege, Testing a multivariate model for wavelet coefficients, in: 2011 18th IEEE International Conference on Image Processing, IEEE, 2011, pp. 1277–1280.
- [39] Mit vision and modeling group, vision texture (95). URL: <https://vismod.media.mit.edu/pub/VisTex/>.
- [40] A. Shiriyayev, *Probability*. 1988, Springer-Verlag (1988).