Adaptive online estimation for mixtures of ECD: a geometric approach

Jialun Zhou, Salem Said, Yannick Berthoumieu

Adaptive online estimation for mixtures of ECD : a geometric approach

Jialun Zhou, Salem Said, and Yannick Berthoumieu

***Abstract*—Mixtures of elliptically-contoured distributions are highly versatile at modeling real-world probability distributions. They have therefore played a valuable role in computer vision and image processing, radar and biomedical signal processing. Existing methods for the estimation of these mixtures may become impractical for relatively-large datasets, either due to lack of computational resources or to poor performance (slow convergence or inaccuracy). To overcome these issues, the present paper introduces a new estimation method, called the CIG method (component-wise information gradient). On the one hand, this is an online method, so it requires moderate computational resources. On the other hand, it uses an adaptive step-size selection rule which guarantees a fast rate of convergence. Based on a geometric approach to the underlying estimation problem, the CIG method derives its name from the introduction of a new information metric on the mixture parameter space, which is called the component-wise information metric, and serves as a substitute for the Fisher information metric.**

***Index Terms*— elliptically-contoured distribution, mixture, information metric, online estimation, texture segmentation**

I. Introduction

Elliptically-contoured distributions (ECD) are a far-reaching generalization of multivariate Gaussian distributions [1]–[3]. An ECD is given by a location vector and a scatter matrix (the mean and covariance, for a multivariate Gaussian), but may also include an additional shape parameter, in order to allow for data with heavy tails or outliers. As such, ECD encompass many widely-used statistical distributions, such as multivariate generalized Gaussian distributions (MGGD) [4], and multivariate Student T-distributions (MSTD) [5]. In turn, mixtures of ECD have been recognized as highly-useful generalizations of Gaussian mixture models [6], [7], which appear in a broad range of applications, such as action recognition, image denoising, robust modeling, clustering and classification of data with outliers [8]–[13], among others [14]–[16].

Estimation of mixture models typically relies on the expectation-maximization (EM) method. Alternatively, for mixtures of Gaussian distributions, a geometric approach, based on Riemannian optimization, was introduced in [17]. This was extended to mixtures of ECD in [18], but under the restriction that shape parameters should be known in advance. For mixtures of MGGD, a comparison between several EM-based methods, and an online estimation method, based on stochastic gradient descent, was carried out in [8]. For mixtures of MSTD, an alternating proximal minimization method (inertial PALM) was applied in [13]. Recently, a geometric approach, based on the Fisher information metric, was applied to mixtures of scaled Gaussian distributions [19].

The above-mentioned methods suffer from certain drawbacks. First, some of them are off-line (*i.e.* batch) methods [18], [19]. These require excessive resources in time and memory for relatively large datasets. Second, when the shape parameter is unknown, existing online methods require mini-batches of increasing size [13]. However, increasing the size of mini-batches may lead back to the problem encountered with off-line methods. Third, these methods either focus on just one subfamily of ECD (MGGD in [8] or MSTD in [13]), or assume the shape parameter is known in advance [18], [19].

The present paper, hoping to overcome these drawbacks, introduces a new online estimation method, called the component-wise information gradient method (CIG). The main features of the CIG method may be summarized as follows

- CIG is an online method, which uses mini-batches of fixed size (one mini-batch per iteration). This reduces the computation time required to perform a single parameter update.
- CIG uses an adaptive step-size selection rule in order to speed up convergence. This reduces the total number of parameter updates required for convergence.
- CIG applies to many widely-used subfamilies of ECD, such as MGGD and MSTD (of course, this does not mean that it applies to multiple subfamilies at once).

The starting point of the CIG method is to formulate the problem of estimating a mixture of ECD as the minimization of a cross-entropy function, defined in Section II. Ideally, one hopes to introduce the Fisher information metric, and apply the corresponding natural gradient method to minimize the cross-entropy [20]. For mixtures of ECD, this information metric does not have a closed form expression. It is therefore replaced by the component-wise information metric, which is introduced in Section III. The gradient of the cross-entropy with respect to this new metric is the component-wise information gradient, the CIG. The CIG method is then presented in Section IV. It is a Riemannian stochastic gradient method, based on the component-wise information gradient. In addition, it implements an adaptive step-size selection, which leads to a fast rate of convergence (when a suitable initialization is used, as stated in Proposition 1). The paper closes with Section V, which compares the CIG method to other state-of-the-art methods, through an application to texture image segmentation.

Jialun Zhou is with the School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China. (e-mail: seiezjl@zzu.edu.cn)

Salem said is with the Laboratoire Jean Kuntzmann, UMR 5224, F-38400, Saint-Martin-d'Hères, France. (e-mail: salem.said@univ-grenoble-alpes.fr)

Yannick Berthoumieu is with the University of Bordeaux, CNRS, IMS, UMR 5218, F-33400, Talence, France. (e-mail: yannick.berthoumieu@u-bordeaux.fr).

II. Formulation of the Problem

An ECD has three parameters, the location parameter μ, the scatter matrix Σ, and the shape parameter β (for an MSTD, β is called the degrees of freedom parameter), which are denoted $\theta = (\mu, \Sigma, \beta)$. If x is a p-dimensional random vector generated from an ECD, with probability density $p(x|\theta)$, then

$$p(x|\theta) = c(\beta)\,[\det(\Sigma)]^{-\frac{1}{2}}\, g\,(\delta, \beta) \tag{1}$$

where $c(\beta)$ is a normalizing factor that depends only on β, and $\delta = (\mu - x)^{\dagger}\Sigma^{-1}(\mu - x)$. The generating function g determines the specific subfamily of ECD. For example,

$$\begin{aligned} g(\delta,\beta) &= \exp\left(-\delta^{\beta}/2\right) \qquad \text{for MGGD} \\ g(\delta,\beta) &= (1+\delta/\beta)^{-\frac{p+\beta}{2}} \qquad \text{for MSTD} \end{aligned} \tag{2}$$

A mixture of ECD is a weighted combination of a finite number K of ECD, which are known as mixture components. The k-th mixture component has density $p(x|\theta_k)$ of the form (1), and is assigned a weight $w_k \in (0,1)$. The mixture is parameterized by $\boldsymbol{\theta} = (w_k, \theta_k; k = 1, \ldots, K)$, and has probability density

$$f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k\, p\,(x; \theta_k) \tag{3}$$

The weights w_k satisfy the normalizing condition $\sum w_k = 1$.

The aim of the present paper is to introduce a new method for online estimation of $\boldsymbol{\theta}$. The starting point of this method is the minimization of the cross-entropy [21],

$$\arg\min_{\boldsymbol{\theta}} D(\boldsymbol{\theta}) \qquad D(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}^*}\left[\ln f\,(x|\boldsymbol{\theta})\right] \tag{4}$$

where $\boldsymbol{\theta}^*$ is the (unknown) true value of the parameter $\boldsymbol{\theta}$. This minimization is carried out by introducing an original Riemannian metric, which will be called the component-wise information metric.

III. The component-wise information metric

To introduce the component-wise information metric, it is convenient to replace the weights w_k with quadratic weights r_k, defined by $w_k = r_k^2$. The normalizing condition on the w_k (stated after (3)) means that the vector $r = (r_k\,; k = 1, \ldots, K)$ always belongs to the unit sphere in $\mathbb{R}^K$, denoted S^{K-1}. Replacing the w_k by the r_k, the mixture density (3) is now parameterized by $\boldsymbol{\theta} = (r, \theta_k\,; k = 1, \ldots, K)$.

Each θ_k belongs to the product space $\Theta = \mathbb{R}^p \times \mathcal{P}_p \times \mathbb{R}_+$, where $\mathcal{P}_p$ denotes the space of $p \times p$ positive-definite matrices. Indeed, $\theta = (\mu, \Sigma, \beta)$ where $\mu \in \mathbb{R}^p$, $\Sigma \in \mathcal{P}_p$, while $\beta \in \mathbb{R}_+$. Therefore, $\boldsymbol{\theta}$ belongs to the product space $\boldsymbol{\Theta} = S^{K-1} \times \Theta^K$, since $r \in S^{K-1}$ and each $\theta_k \in \Theta$ (for $k = 1, \ldots, K$).

Ideally, one hopes to equip $\boldsymbol{\Theta}$ with the Fisher information metric. However, this information metric does not have a closed form expression. As an alternative, the component-wise information metric is here introduced,

$$\langle U, V\rangle_{\boldsymbol{\theta}} = \langle U^r, V^r\rangle_r + \sum_{k=1}^{K} \langle U^{\theta_k}, V^{\theta_k}\rangle_{\theta_k} \tag{5}$$

where U, V are tangent vectors to $\boldsymbol{\Theta}$ in the tangent space $T_{\boldsymbol{\theta}}\boldsymbol{\Theta}$, $U = (U^r, U^{\theta_k}; k = 1, \ldots, K)$ and similarly $V = (V^r, V^{\theta_k}; k = 1, \ldots, K)$.

In (5), $\langle \cdot, \cdot \rangle_r$ denotes the usual scalar product in the Euclidean space $\mathbb{R}^K$. On the other hand, for each $\theta_k = (\mu_k, \Sigma_k, \beta_k)$, one has the information metric given in [22],

$$\langle U^{\theta_k}, V^{\theta_k}\rangle_{\theta_k} = \langle U^{\mu_k}, V^{\mu_k}\rangle_{\mu_k} + \langle U^{\Sigma_k}, V^{\Sigma_k}\rangle_{\Sigma_k} + \langle U^{\beta_k}, V^{\beta_k}\rangle_{\beta_k}$$

Here, if $U^{\theta} = (U^{\mu}, U^{\Sigma}, U^{\beta})$ and $V^{\theta} = (V^{\mu}, V^{\Sigma}, V^{\beta})$, then

$$\begin{aligned} \langle U^{\mu}, V^{\mu}\rangle_{\mu} &= I_{\mu}\, U^{\mu\dagger}\, \Sigma^{-1}\, V^{\mu} \\ \langle U^{\Sigma}, V^{\Sigma}\rangle_{\Sigma} &= I_{\Sigma,1}\mathrm{tr}\left(\Sigma^{-1}U^{\Sigma}\Sigma^{-1}V^{\Sigma}\right) \\ &\qquad + I_{\Sigma,2}\mathrm{tr}\left(\Sigma^{-1}U^{\Sigma}\right)\mathrm{tr}\left(\Sigma^{-1}V^{\Sigma}\right) \\ \langle U^{\beta}, V^{\beta}\rangle_{\beta} &= I_{\beta} U^{\beta} V^{\beta} \end{aligned} \tag{6}$$

in terms of the information constants $(I_{\mu}, I_{\Sigma,1}, I_{\Sigma,2}, I_{\beta})$, which are given explicitly for MGGD and MSTD in the supplementary material.

Based on the component-wise information metric (5), the component-wise information gradient (CIG) may now be introduced. If $\mathfrak{q} = -\ln f\,(x|\boldsymbol{\theta})$ (f was defined in (3)), then the CIG is the unique vector field $\nabla^{\odot}\mathfrak{q}$ on $\boldsymbol{\Theta}$ which satisfies [23]

$$\left\langle \nabla^{\odot}\mathfrak{q}\,(\boldsymbol{\theta}; x)\,, V\right\rangle_{\boldsymbol{\theta}} = \mathrm{d}\mathfrak{q}\,(\boldsymbol{\theta}; x)\cdot V \tag{7}$$

for any tangent vector $V \in T_{\boldsymbol{\theta}}\boldsymbol{\Theta}$, where $\mathrm{d}\mathfrak{q}(x|\boldsymbol{\theta})$ denotes the differential of the negative log-likelihood, $\mathfrak{q}(x|\boldsymbol{\theta})$.

The expectation of the CIG is the component-wise gradient of the cross-entropy $D(\boldsymbol{\theta})$, which was defined in (4),

$$\mathbb{E}_{\boldsymbol{\theta}^*}\left[\nabla^{\odot}\mathfrak{q}\,(\boldsymbol{\theta}; x)\right] = \nabla^{\odot} D(\boldsymbol{\theta}) \tag{8}$$

Moreover, the CIG has a component-wise structure,

$$\nabla^{\odot}\mathfrak{q}\,(\boldsymbol{\theta}; x) = \begin{pmatrix} \nabla_r \mathfrak{q}\,(\boldsymbol{\theta}; x) \\ \left(\nabla_{\theta_k}\mathfrak{q}\,(\boldsymbol{\theta}; x)\right)_{k=1,\cdots,K} \end{pmatrix}$$

where

$$\nabla_{\theta}\mathfrak{q}\,(\boldsymbol{\theta}; x) = \begin{pmatrix} \left(\nabla_{\mu}\mathfrak{q}\,(\boldsymbol{\theta}; x)\right) \\ \left(\nabla_{\Sigma}\mathfrak{q}\,(\boldsymbol{\theta}; x)\right) \\ \left(\nabla_{\beta}\mathfrak{q}\,(\boldsymbol{\theta}; x)\right) \end{pmatrix}$$

and the individual components are given as follows. First,

$$\nabla_r \mathfrak{q}\,(\boldsymbol{\theta}; x) = \partial\mathfrak{q}(\boldsymbol{\theta}; x)/\partial r - \langle \partial\mathfrak{q}(\boldsymbol{\theta}; x)/\partial r, r\rangle_r \times r \tag{9}$$

where $\langle \cdot, \cdot \rangle_r$ denotes the scalar product in $\mathbb{R}^K$, as in (5). Second, for the remaining components,

$$\begin{aligned} \nabla_{\mu}\mathfrak{q}(\boldsymbol{\theta}; x) &= I_{\mu}^{-1}\,\Sigma\,\mathrm{G}_{\mu}(\boldsymbol{\theta}, x) \\ \nabla_{\Sigma}\mathfrak{q}(\boldsymbol{\theta}; x) &= J_{\Sigma,1}\mathrm{G}_{\Sigma}^{\perp}(\boldsymbol{\theta}, x) + J_{\Sigma,2}\mathrm{G}_{\Sigma}^{\parallel}(\boldsymbol{\theta}, x) \\ \nabla_{\beta}\mathfrak{q}(\boldsymbol{\theta}; x) &= I_{\beta}^{-1}\,\mathrm{G}_{\beta}(\boldsymbol{\theta}, x) \end{aligned} \tag{10}$$

where the coefficients $(J_{\Sigma,1},\, J_{\Sigma,2})$ are

$$J_{\Sigma,1} = 1/I_{\Sigma,1} \qquad J_{\Sigma,2} = 1/(I_{\Sigma,1} + p\, I_{\Sigma,2}) \tag{11}$$

in terms of the information constants $(I_{\Sigma,1}, I_{\Sigma,2})$ from (6), and where $\mathrm{G}_{\mu}(\boldsymbol{\theta}, x)$, $\mathrm{G}_{\beta}(\boldsymbol{\theta}, x)$ are Euclidean gradients, while $\mathrm{G}_{\Sigma}(\boldsymbol{\theta}, x)$ is the affine-invariant gradient [24]. Moreover, $\perp$ and $\parallel$ denote the parallel and orthogonal components of $\mathrm{G}_{\Sigma}(\boldsymbol{\theta}, x)$, as defined in [25],

$$\begin{aligned} \mathrm{G}_{\Sigma}^{\parallel}(\boldsymbol{\theta}, x) &= \frac{1}{p}\mathrm{tr}\left(\Sigma^{-1}\mathrm{G}_{\Sigma}(\boldsymbol{\theta}, x)\right)\Sigma \\ \mathrm{G}_{\Sigma}^{\perp}(\boldsymbol{\theta}, x) &= \mathrm{G}_{\Sigma}(\boldsymbol{\theta}, x) - \mathrm{G}_{\Sigma}^{\parallel}(\boldsymbol{\theta}, x) \end{aligned} \tag{12}$$

Detailed expressions of $\mathrm{G}_{\mu}(\boldsymbol{\theta}, x)$, $\mathrm{G}_{\Sigma}(\boldsymbol{\theta}, x)$ and $\mathrm{G}_{\beta}(\boldsymbol{\theta}, x)$ are given in the supplementary material.

IV. The CIG Method

The main contribution of the present paper is to introduce the CIG method. This is an online estimation method which uses an adaptive step-size in order to speed up convergence. Its basic idea is to use the CIG $\nabla^{\odot}\mathfrak{q}$, defined in the previous section, in order to search for the minimum in (4).

Starting from an initialization $\boldsymbol{\theta}^{(0)}$, the CIG method uses a stream of mini-batches $(B^{(t)}; t = 1, 2, \ldots)$, in order to compute a sequence of estimates $\boldsymbol{\theta}^{(t)}$ which approximate the true parameter $\boldsymbol{\theta}^*$, under suitable conditions (see Proposition 1, below). The update from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$ only uses the new mini-batch $B^{(t+1)}$. This provides the mini-batch CIG,

$$\nabla^{\odot}\mathfrak{q}(\boldsymbol{\theta}^{(t)}|B^{(t+1)}) = \frac{1}{|B^{(t+1)}|}\sum_{x\in B^{(t+1)}} \nabla^{\odot}\mathfrak{q}\left(\boldsymbol{\theta}; x\right) \qquad (13)$$

where the sum is over samples x which belong to the new mini-batch $B^{(t+1)}$, whose size is here denoted by $|B^{(t+1)}|$. This mini-batch CIG has a component-wise structure, due to the component-wise structure of $\nabla^{\odot}\mathfrak{q}\left(\boldsymbol{\theta}; x\right)$, given by (9)-(10). The resulting components of $\nabla^{\odot}\mathfrak{q}(\boldsymbol{\theta}^{(t)}|B^{(t+1)})$ are denoted by $\nabla_r\mathfrak{q}\left(\boldsymbol{\theta}; B^{(t+1)}\right)$, $\nabla_\mu\mathfrak{q}(\boldsymbol{\theta}; B^{(t+1)})$, $\nabla_\Sigma\mathfrak{q}(\boldsymbol{\theta}; B^{(t+1)})$, $\nabla_\beta\mathfrak{q}(\boldsymbol{\theta}; B^{(t+1)})$, as computed from (9)-(10) and (13).

Accordingly, $\boldsymbol{\theta}^{(t)} = (r^{(t)}, \theta_k^{(t)}; k = 1, \ldots, K)$ is updated to obtain $\boldsymbol{\theta}^{(t+1)} = (r^{(t+1)}, \theta_k^{(t+1)}; k = 1, \ldots, K)$, as follows. First,

$$r^{(t+1)} = \mathrm{Exp}_{r^{(t)}}\left(-\eta^{(t+1)}\nabla_r\mathfrak{q}\left(\boldsymbol{\theta}^{(t)}|B^{(t+1)}\right)\right) \qquad (14)$$

where $\eta^{(t+1)}$ is the adaptive step-size (defined in (18), below) and Exp is the Riemannian exponential on the sphere S^{K-1},

$$\mathrm{Exp}_r\left(V^r\right) = \cos\left(\|V^r\|\right) r + \sin\left(\|V^r\|\right)\frac{V^r}{\|V^r\|}$$

for any $r \in S^{K-1}$ and tangent vector $V^r \in T_r S^{K-1}$. Second, for the remaining parameters $\theta_k^{(t)} = (\mu_k^{(t)}, \Sigma_k^{(t)}, \beta_k^{(t)})$,

$$\mu_k^{(t+1)} = \mu_k^{(t)} - \eta^{(t+1)}\nabla_{\mu_k}\mathfrak{q}\left(\boldsymbol{\theta}^{(t)}|B^{(t+1)}\right) \qquad (15)$$

Moreover,

$$\Sigma_k^{(t+1)} = \mathrm{Exp}_{\Sigma_k^{(t)}}\left(-\eta^{(t+1)}\nabla_{\Sigma_k}\mathfrak{q}\left(\boldsymbol{\theta}^{(t)}|B^{(t+1)}\right)\right) \qquad (16)$$

where Exp is the Riemannian exponential on $\mathcal{P}_p$ (here, exp denotes the matrix exponential [24]),

$$\mathrm{Exp}_\Sigma(V^\Sigma) = \Sigma\exp\left(\Sigma^{-1}V^\Sigma\right)$$

for any $\Sigma \in \mathcal{P}_p$ and tangent vector $V^\Sigma \in T_\Sigma\mathcal{P}_p$. Furthermore,

$$\beta_k^{(t+1)} = \mathrm{Exp}_{\beta_k^{(t)}}\left(-\eta^{(t+1)}\nabla_{\beta_k}\mathfrak{q}\left(\boldsymbol{\theta}^{(t)}|B^{(t+1)}\right)\right) \qquad (17)$$

with the notation

$$\mathrm{Exp}_\beta(V^\beta) = \beta \times e^{V^\beta/\beta}$$

for $\beta > 0$ and $V^\beta \in \mathbb{R}$. The CIG method repeats the updates (14)-(17), whenever a new mini-batch becomes available.

The method has two essential features. First, it is an online method, where the update from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$ only relies on $B^{(t+1)}$. Second, it uses an adaptive step-size $\eta^{(t+1)}$, which is computed based on the current $\boldsymbol{\theta}^{(t)}$. Specifically,

$$\eta^{(t+1)} = \frac{\tau_{min}^{(t)}}{L\,\tau_{max}^{(t)}} \qquad (18)$$

Here, L is a certain constant, which will shortly be introduced in Proposition 1, and $\tau_{min}^{(t)}$ and $\tau_{max}^{(t)}$ are given by

$$\begin{aligned}\tau_{min}^{(t)} &= \min\left\{1, \left(\lambda_{min,k}^{(t)}/I_{\mu_k}^{(t)}\right)_k, \left(J_{\Sigma_k,2}^{(t)}\right)_k, \left(1/I_{\beta_k}^{(t)}\right)_k\right\}\\ \tau_{max}^{(t)} &= \max\left\{1, \left(\lambda_{max,k}^{(t)}/I_{\mu_k}^{(t)}\right)_k^2, \left(J_{\Sigma_k,1}^{(t)}\right)_k^2, \left(1/I_{\beta_k}^{(t)}\right)_k^2\right\}\end{aligned} \qquad (19)$$

where $\lambda_{min,k}$ and $\lambda_{max,k}$ are the smallest and largest eigenvalues of Σ_k, and where the information constants ($I_{\mu_k}^{(t)}$, $J_{\Sigma_k,2}^{(t)}$, *etc.*) are computed as in (6) and (10), based on the current value of the parameters $(r^{(t)}, \mu_k^{(t)}, \Sigma_k^{(t)}, \beta_k^{(t)})$.

The rate of convergence and asymptotic behavior of the CIG method are described in the following Proposition 1, roughly based on [26]. The precise statement of conditions required in this Proposition, as well as a sketch of its proof, may be found in the supplementary material.

Proposition 1 (fast convergence of CIG): There exists a compact and convex neighborhood $\boldsymbol{\Theta}^*$ of $\boldsymbol{\theta}^*$ in which the objective function $D(\boldsymbol{\theta})$ satisfies the following conditions

(i) it is L-geodesically smooth (see Equation (28) in the supplementary), for some $L > 0$.
(ii) it is α-geodesically strongly convex (see Equation (33) in the supplementary), for some $\alpha > 0$.

Moreover, if $(\boldsymbol{\theta}^{(t)})$, generated by the CIG method (14)–(17) with adaptive step-size (18) and constant mini-batch size b, remains within $\boldsymbol{\Theta}^*$, then it achieves a fast rate of convergence

$$\mathbb{E}_{\boldsymbol{\theta}^*}\left[D\left(\boldsymbol{\theta}^{(T)}\right) - D\left(\boldsymbol{\theta}^*\right)\right] \leqslant c_T\left[D\left(\boldsymbol{\theta}^{(0)}\right) - D\left(\boldsymbol{\theta}^*\right)\right] + \varepsilon_*^2 \qquad (20)$$

where $\varepsilon_*^2 > 0$ is a constant and where

$$c_T = \prod_{t=0}^{T}\left\{1 - \frac{\alpha\left[\tau_{min}^{(t)}\right]^2}{L\,\tau_{max}^{(t)}}\right\} \qquad (21)$$

The rate of convergence (20) is called fast, because it is geometric. Precisely, it can be shown, based on elementary arguments, that there exists some $c \in (0, 1)$ such that $c_T \leq c^T$.

With this geometric rate, the difference between $\boldsymbol{\theta}^{(T)}$ and $\boldsymbol{\theta}^{(*)}$ (measured by the left-hand side of (20)) converges to a limit smaller than the positive constant ε_*^2, which quantifies the asymptotic accuracy of the method. It should be noted ε_*^2 is proportional to $1/b$ (where b is the mini-batch size).

In practice, when implementing the adaptive step-size (18), large values of L lead to slow convergence at early stages of the estimation. To overcome this issue, the following step-size may be used,

$$\eta^{(t+1)} = \max\left\{\frac{\tau_{min}^{(t)}}{L\,\tau_{max}^{(t)}}, \frac{a}{t+1}\right\} \qquad (22)$$

a combination of (18) and of the decreasing step-size used in [22], which involves a constant a, selected as in [22], [25]. Moreover, for simple scenarios, with small dimension p or number of components K (recall the notation of Section II), it is possible to replace the adaptive step-size (18) with a constant step-size $< 1/L$, and still obtain similar results. However, this yields poor performance when $K > 2$, and will always requires manual selection of the constant step-size.

The following section will present an application of the CIG method to texture segmentation. This will highlight the main features of the CIG method, and compare it to state-of-the-art online estimation methods.

V. Texture segmentation

The CIG method was applied to texture segmentation, within images from the DTD database [27]. Specifically, textures were modeled using mixtures of MGGD and MSTD, motivated by the fact that these two subfamilies of ECD successfully capture the wavelet statistics of texture images, as shown in [9], [28], [29].

Two images were randomly selected from the DTD database, 'honeycombed' and 'pitted'. A mixed image was then considered in 3-dimensional CIE-Lab color space (see Figure 1). Since the two images have similar texture and color features, the two resulting point clouds in color space are quite hard to separate (see Figure 1(c)).

Segmentation was carried out using a classical Bayesian classifier [30]. In other words, each single pixel was identified as 'honeycombed' or 'pitted', based on the maximum a posteriori rule. Recall from Bayes formula that

$$a\ posteriori = (prior \times likelihood)/evidence \qquad (23)$$

The evidence is just a normalizing factor, and plays no role in the following.

The likelihood was chosen to be a mixture of MGGD or MSTD with $K = 2$ mixture components. This choice reflects the fact that the two point clouds (blue and orange in Figure 1(c)) do not appear to have an ellipsoidal shape, and therefore cannot be represented by single MGGD or MSTD. The prior probabilities are the overall probabilities of being 'honeycombed' or 'pitted' (blue or orange). They were estimated empirically, by a straightforward uniform sampling.

The CIG method, and three other state-of-the-art online estimation methods, were used to estimate the likelihood distributions. These three methods are Euclidean stochastic gradient (SGD [8]) and affine-invariant stochastic gradient (AIG [17]), in addition to the general Adam method, widely regarded as one of the most robust online optimization methods [31].

Each of the two original images has more than 5×10^5 pixels. Half of these were used as training set, and the remaining half made up the mixed image of Figure 1, used as the test set.

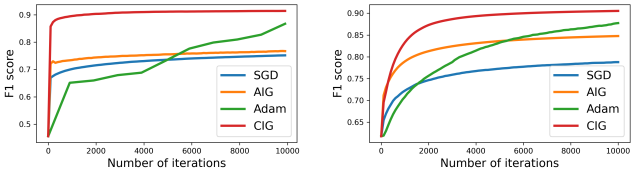
The same randomly-chosen initialization, and the same mini-batch size ($= 10$), were used in each of the four above-mentioned methods (CIG, SGD, AIG, Adam). SGD and AIG were implemented with a decreasing step-size, while Adam was implemented with the standard parametrization proposed in [31]. For CIG, the step-size in (22) was used with $a = 1$.

The four methods were evaluated by the F1 score (defined in [32]), as shown in Figure 2. The CIG method has a significant advantage in terms of speed of convergence and asymptotic F1 score (after convergence).

The final segmentation results are presented in Figure 3. CIG is visibly better than the other three methods, since the right-most column of Figure 3 has the clearest separation between light and dark gray. Table I reports mean and standard deviation of the computation time per one iteration (in milliseconds). This table shows no significant difference between the four methods.

(a) Mixed image (b) True seg. (c) CIE-Lab

Fig. 1: Mixed texture image



(a) MGGD (b) MSTD

Fig. 2: Number of iteration versus F1 score

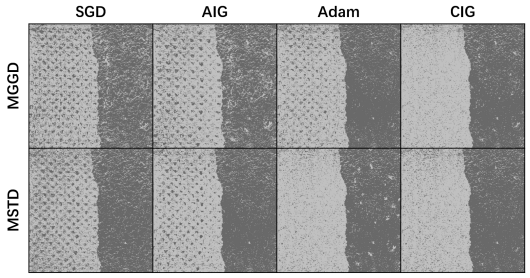


Fig. 3: Segmentation results

TABLE I: Time per iteration (ms): mean/standard deviation

methods	SGD	AIG	Adam	CIG
mean	0.9	1	0.9	1
s.d.	$3.1 \cdot 10^{-2}$	$3.3 \cdot 10^{-3}$	$3.2 \cdot 10^{-2}$	$1.2 \cdot 10^{-3}$

VI. Conclusion

The present paper has introduced the CIG method for the estimation of mixtures of ECD. This is an online estimation method, which uses an adaptive step-size in order to speed up convergence. Theoretically, it was shown that the CIG method, when initialized correctly, guarantees a fast (precisely, geometric) rate of convergence. Numerically, the CIG method was applied to texture segmentation, with images taken from the DTD database. This application showcased the advantages of this CIG method, in terms of its rate of convergence and asymptotic accuracy, compared to state-of-the-art online estimation methods. In fact, the CIG method requires roughly the same amount of time, as other methods, to perform a single iteration, while achieving convergence after a significantly smaller number of iterations. Moreover, after convergence, the CIG method was seen to provide a significantly greater accuracy, with regard to the texture segmentation problem at hand. Although the CIG method was only applied to mixtures of MGGD and MSTD, it promises to extend successfully to several other subfamilies of ECD.

REFERENCES

[1] D. Kelker, "Distribution theory of spherical distributions and a location-scale parameter generalization," Sankhyā: The Indian Journal of Statistics, Series A, pp. 419–430, 1970.

[2] K. Fang and 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-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.

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

[6] H. Holzmann, A. Munk, and T. Gneiting, "Identifiability of finite mixtures of elliptical distributions," Scandinavian journal of statistics, vol. 33, no. 4, pp. 753–763, 2006.

[7] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," IEEE Transactions on pattern analysis and machine intelligence, vol. 24, no. 3, pp. 381–396, 2002.

[8] F. Najar, S. Bourouis, R. Al-Azawi, and A. Al-Badi, "Online recognition via a finite mixture of multivariate generalized Gaussian distributions," in Mixture Models and Applications. Springer, 2020, pp. 81–106.

[9] S. Tan and L. Jiao, "Multivariate statistical models for image denoising in the wavelet domain," International Journal of Computer Vision, vol. 75, no. 2, pp. 209–230, 2007.

[10] D. Peel and G. J. McLachlan, "Robust mixture modelling using the T distribution," Statistics and computing, vol. 10, no. 4, pp. 339–348, 2000.

[11] J. L. Andrews and P. D. McNicholas, "Model-based clustering, classification, and discriminant analysis via mixtures of multivariate T-distributions," Statistics and Computing, vol. 22, no. 5, pp. 1021–1029, 2012.

[12] T.-I. Lin, P. D. McNicholas, and H. J. Ho, "Capturing patterns via parsimonious T mixture models," Statistics & Probability Letters, vol. 88, pp. 80–87, 2014.

[13] J. Hertrich and G. Steidl, "Inertial stochastic PALM and applications in machine learning," Sampling Theory, Signal Processing, and Data Analysis, vol. 20, no. 1, pp. 1–33, 2022.

[14] B. Ge, N. Bouguila, and W. Fan, "Single-target visual tracking using color compression and spatially weighted generalized Gaussian mixture models," Pattern Analysis and Applications, vol. 25, no. 2, pp. 285–304, 2022.

[15] D. Yapi and M. S. Allili, "Multi-band texture modeling using finite mixtures of multivariate generalized Gaussian distributions," in 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022, pp. 464–469.

[16] D.-P.-L. Nguyen, J.-F. Aujol, and Y. Berthoumieu, "Patch-based image super resolution using generalized Gaussian mixture model," arXiv preprint arXiv:2206.03069, 2022.

[17] R. Hosseini and S. Sra, "An alternative to EM for Gaussian mixture models: batch and stochastic Riemannian optimization," Mathematical programming, vol. 181, no. 1, pp. 187–223, 2020.

[18] S. Li, Z. Yu, and D. Mandic, "A universal framework for learning the elliptical mixture model," IEEE Transactions on Neural Networks and Learning Systems, 2020.

[19] A. Collas, "Riemannian geometry for statistical estimation and learning: application to remote sensing," Ph.D. dissertation, université Paris-Saclay, 2022.

[20] S.-I. Amari, "Natural gradient works efficiently in learning," Neural computation, vol. 10, no. 2, pp. 251–276, 1998.

[21] T. Cover and J. Thomas, Elements of Information Theory. Wiley, 2012.

[22] J. Zhou, S. Said, and Y. Berthoumieu, "Riemannian information gradient methods for the parameter estimation of ECD," Signal Processing, vol. 192, p. 108376, 2022.

[23] P.-A. Absil, R. Mahony, and R. Sepulchre, Optimization algorithms on matrix manifolds. Princeton University Press, 2009.

[24] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian Framework for Tensor Computing," INRIA, Tech. Rep. RR-5255, Jul. 2004.

[25] J. Zhou and S. Said, "Fast, asymptotically efficient, recursive estimation in a Riemannian manifold," Entropy, vol. 21, no. 10, p. 1021, 2019.

[26] S. Y. Meng, S. Vaswani, I. H. Laradji, M. Schmidt, and S. Lacoste-Julien, "Fast and furious convergence: Stochastic second order methods under interpolation," in International Conference on Artificial Intelligence and Statistics. PMLR, 2020, pp. 1375–1386.

[27] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi, "Describing Textures in the Wild," in Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2014.

[28] G. Verdoolaege and P. Scheunders, "Geodesics on the manifold of multivariate generalized Gaussian distributions with an application to multicomponent texture discrimination," International Journal of Computer Vision, vol. 95, no. 3, pp. 265–286, 2011.

[29] R. Kwitt, P. Meerwald, A. Uhl, and G. Verdoolaege, "Testing a multivariate model for wavelet coefficients," in 2011 18th IEEE International Conference on Image Processing. IEEE, 2011, pp. 1277–1280.

[30] S. J. Russell, Artificial intelligence a modern approach. Pearson Education, Inc., 2010.

[31] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," Computer Science, 2014.

[32] C. Van Rijsbergen and C. Van Rijsbergen, Information Retrieval. Butterworths, 1979.

[33] O. Besson and I. Abramovich, Yuri, "On the Fisher information matrix for multivariate elliptically contoured distributions," IEEE Signal Processing Letters, vol. 20, no. 11, pp. 1130–1133, 2013.

[34] G. Verdoolaege and P. Scheunders, "On the geometry of multivariate generalized Gaussian models," Journal of Mathematical Imaging and Vision, vol. 43, pp. 180–193, 2012.

SUPPLEMENTARY MATERIAL

To begin, the expressions of the information constants and of the Euclidean and affine-invariant gradients, appearing in (6) and (10), are here provided.

Information constants

For MGGD and MSTD, the information constants appearing in (6) and (10) are found from the following table [33], [34]

	Model	
	MGGD	MSTD
I_μ	$\frac{(p-2)[2(\beta-1)+p]\Gamma(\frac{p-2}{2\beta})}{p2^{\frac{1}{\beta}}\Gamma(\frac{p}{2\beta})}$	$\frac{p+\beta}{p+\beta+2}$
$J_{\Sigma,1}$	$\frac{2(p+2)}{p+2\beta}$	$\frac{2(p+\beta+2)}{p+\beta}$
$J_{\Sigma,2}$	$\frac{2}{\beta}$	$\frac{2(p+\beta+2)}{\beta}$
I_β	$\frac{1+P_1+P_2+P_3}{\beta^2}$	$\frac{Q_1-p(p+\beta+4)}{2\beta(p+\beta)(p+\beta+2)}$

TABLE II: Information constants

and from the relations

$$J_{\Sigma,1} = I_{\Sigma,1}^{-1} \qquad J_{\Sigma,2} = 1/\left(I_{\Sigma,1} + pI_{\Sigma,2}\right) \tag{24}$$

In the table,

$$\begin{aligned} P_1 &= (p/2\beta)^2\,\Psi_1(p/2\beta) \\ P_2 &= (p/\beta)\left[\ln 2 + \Psi_0(p/2\beta)\right] \\ P_3 &= (p/2\beta)\left\{P_4 + \Psi_1(1+p/2\beta)\right\} \\ P_4 &= (\ln 2)^2 + P_5\left[\ln 4 + P_5\right] \\ P_5 &= \Psi_0(1+p/2\beta) \\ Q_1 &= \left\{\Psi_1(\beta/2) - \Psi_1[(p+\beta)/2]\right\}/4 \end{aligned} \tag{25}$$

where Ψ_0 and Ψ_1 denote the polygamma functions.

Gradients

The Euclidean and affine-invariant gradients in (10) are [22]

$$\begin{aligned} \mathrm{G}_{\mu_k}(\boldsymbol{\theta}, x) &= 2o_k \frac{\partial h(\delta_k, x)}{\partial \delta_k}\Sigma^{-1}(x-\mu_k) \\ \mathrm{G}_{\Sigma_k}(\boldsymbol{\theta}, x) &= o_k\left[\frac{1}{2}\Sigma_k + \frac{\partial h(\delta_k,\beta_k)}{\partial \delta_k}S_k\right] \\ \mathrm{G}_{\beta_k}(\boldsymbol{\theta}, x) &= -o_k\left[\frac{\partial \ln c(\beta_k)}{\partial \beta_k} + \frac{\partial h(\delta_k,\beta_k)}{\partial \beta_k}\right] \end{aligned} \tag{26}$$

where $h = \ln(g)$, with g the generating function from (1), $S_k = (x-\mu_k)(x-\mu_k)^\dagger$ and $o_k = w_k p(x|\theta_k)/\sum_{j=1}^K w_j p(x|\theta_j)$.

A. Sketch of proof

Note Θ^* is the neighborhood of $\boldsymbol{\theta}^*$ which satisfies all assumptions in proposition 1. Here, another product metric is used for proving Proposition 1, which is constructed by the sum of classic intrinsic Riemannian metrics within each sub spaces

$$\begin{aligned} \tilde{g}(u,v) = \langle U^r, V^r\rangle + \sum_k \langle U^{\mu_k}, V^{\mu_k}\rangle \\ + \sum_k g(U^{\Sigma_k}, V^{\Sigma_k}) + \sum_k U^{\beta_k}V^{\beta_k} \end{aligned} \tag{27}$$

Where $\langle\cdot,\cdot\rangle$ is dot product in Euclidean space, and $g(\cdot,\cdot)$ is the affine-invariant metric in SPD matrix space. Recall the geodesically L-lipschitz smooth of the cost function $D(\boldsymbol{\theta})$

$$\begin{aligned} D(\boldsymbol{\theta}^{(t+1)}) \leqslant & D(\boldsymbol{\theta}^{(t)}) - \eta^{(t+1)}\tilde{g}\left(\mathrm{G}(\boldsymbol{\theta}^{(t)}), U(\boldsymbol{\theta}^{(t)}, B^{(t+1)})\right) \\ & + \frac{L}{2}\mathrm{d}_{\tilde{g}}^2\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}\right) \end{aligned} \tag{28}$$

where $\mathrm{G}(\boldsymbol{\theta}^{(t)})$ and $\mathrm{d}_{\tilde{g}}(\cdot,\cdot)$ denote the product gradient and distance derived by the product metric (27), respectively. More specifically, the arc length on unit sphere locates in the component of weights, then Eulidean distance and affine invariant distance follow in sequence. The vector $U(\boldsymbol{\theta}^{(t)}, B^{(t)})$ is the descending direction (the CIG), and $B^{(t)}$ is a randomly selected mini-batch. Take expectation of the two sides of equation (28), w.r.t the law for selection of mini-batch $B^{(t+1)}$.

$$\begin{aligned} \mathbb{E}\left[D(\boldsymbol{\theta}^{(t+1)})\right] \leqslant & D(\boldsymbol{\theta}^{(t)}) - \eta^{(t+1)}\tilde{g}\left(\mathrm{G}(\boldsymbol{\theta}^{(t)}), U(\boldsymbol{\theta}^{(t)})\right) \\ & + \frac{L}{2}\mathbb{E}\left[\mathrm{d}_{\tilde{g}}^2\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}\right)\right] \end{aligned} \tag{29}$$

The CIG can be decomposed in tangent space of $\boldsymbol{\theta}^{(t)}$ and represented by the classic gradient. Applying this truth in equation (29), the following equations could be obtained

$$\begin{aligned} \mathrm{d}_{\tilde{g}}^2(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) &= [\eta^{(t+1)}]^2\|U(\boldsymbol{\theta}^{(t)}, B^{(t+1)})\|_{\tilde{g}}^2 \\ \|U(\boldsymbol{\theta}^{(t)}, B^{(t+1)})\|_{\tilde{g}}^2 &\leqslant \tau_{\max}\|\mathrm{G}(\boldsymbol{\theta}^{(t)}, B^{(t)})\|_{\tilde{g}}^2 \\ \tilde{g}\left(\mathrm{G}(\boldsymbol{\theta}^{(t)}), U(\boldsymbol{\theta}^{(t)})\right) &\geqslant \tau_{\min}\|\mathrm{G}(\boldsymbol{\theta}^{(t)})\|_{\tilde{g}}^2 \end{aligned} \tag{30}$$

the coefficients τ_{min} and τ_{max} are given in equation (18) of main article. For any new mini-batch B

$$\mathbb{E}[\|G(\boldsymbol{\theta}, B)\|^2] = \|\mathbb{E}[G(\boldsymbol{\theta}, B)]\|^2 + Var[\|G(\boldsymbol{\theta}, B)\|] \tag{31}$$

In neighborhood of $\boldsymbol{\theta}^*$, the variance $Var[\|G(\boldsymbol{\theta}, B)\|]$ is bounded by some finite positive constant v_*^2

$$\mathbb{E}[\|G(\boldsymbol{\theta}, B)\|^2] \leqslant \|\mathbb{E}[G(\boldsymbol{\theta}, B)]\|^2 + v_*^2 \tag{32}$$

Applying equations (30)-(32) and the step-size $\frac{\tau_{min}^{(t)}}{L\tau_{max}^{(t)}}$ to equation (29), we can obtain

$$\mathbb{E}\left[D\left(\boldsymbol{\theta}^{(t+1)}\right)\right] \leqslant D\left(\boldsymbol{\theta}^{(t)}\right) - \frac{\left[\tau_{min}^{(t)}\right]^2}{2\,L\,\tau_{max}^{(t)}}\left\|\mathrm{G}\left(\boldsymbol{\theta}^{(t)}\right)\right\|_{\tilde{g}}^2 + [\varepsilon_*^{(t)}]^2$$

where $[\varepsilon_*^{(t)}]^2 = \frac{[\tau_{min}^{(t)}]^2 v_*^2}{2L\tau_{max}}$. Finally, using the Polyak-Lojasiewicz inequality

$$\|G(\boldsymbol{\theta}^{(t)})\|_{\tilde{g}} \geqslant 2\alpha(D(\boldsymbol{\theta}^{(t)}) - D(\boldsymbol{\theta}^*)) \tag{33}$$

we can obtain

$$\begin{aligned} & \mathbb{E}\left[D\left(\boldsymbol{\theta}^{(t+1)}\right)\right] - D\left(\boldsymbol{\theta}^*\right) \\ \leqslant & \left(1 - \frac{\left[\tau_{min}^{(t)}\right]^2\alpha}{2\,L\,\tau_{max}^{(t)}}\right)\left[D\left(\boldsymbol{\theta}^{(t)}\right) - D\left(\boldsymbol{\theta}^*\right)\right] + [\varepsilon_*^{(t)}]^{(2)} \end{aligned}$$

Range these T times of iteration

$$\mathbb{E}\left[D\left(\boldsymbol{\theta}^{(T)}\right)\right] - D\left(\boldsymbol{\theta}^*\right) \leqslant c_T\left[D\left(\boldsymbol{\theta}^{(0)}\right) - D\left(\boldsymbol{\theta}^*\right)\right] + \varepsilon_*^2 \tag{34}$$

with

$$c_T = \prod_{t=0}^{T}\left(1 - \frac{\alpha\left[\tau_{min}^{(t)}\right]^2}{2\,L\,\tau_{max}^{(t)}}\right) \tag{35}$$

and with ε_*^2 a positive constant.