Recursive algorithm for estimation of mixture models on Riemannian symmetric space

ZHOU JIALUN[1], LE BIHAN NICOLAS[2], SAID SALEM[1]

[1]Laboratoire IMS
351 Cours de la libération, 33405 Talence cedex, France

[2]CNRS / Gipsa-Lab
11 Rue des mathematiques, 38402 St Martin d'Heres, FRANCE

Jialun.zhou@u-bordeaux.fr, Nicolas.Le-Bihan@gipsa-lab.grenoble-inp.fr,
salem.said@ims-bordeaux.fr

Résumé – Dans de nombreux domaines tels que l'imagerie médicale, la vision par ordinateur et le traitement du signal radar, on est conduit à étudier des distributions de mélange dans l'espace symétrique Riemannien. Cet article propose un algorithme récursive permettant d'estimer les paramètres et de sélectionner simultanément le nombre de composantes d'un modèle de mélange dans un espace symétrique Riemannien. L'idée génératrice est d'initialiser le processus d'estimation à partir d'un grand nombre de composantes K_0, et d'introduire une distribution a priori du poids des composantes pour exprimer notre préférence pour des modèles compacts. En utilisant le gradient d'information de Rao-Fisher pour mettre à jour les paramètres, dans chaque itération, la distribution a priori conduit à l'extinction des composantes non pertinentes. Cet algorithme pourrait être appliqué pour estimer les modèles sur l'espace symétrique. Cet algorithme est simple à utiliser, en effet il est robust par rapport au choix des valeurs initiales, il peut en outre sélectionner le nombre de composantes automatiquement. Nous illustrons cet article par quelques expérimentations.

Abstract – In many fields such as medical imaging, computer vision and radar signal processing, we are lead to study mixtures of distributions in the Riemannian symmetric space. This paper proposes a recursive algorithm for estimating parameters and simultaneously selecting the number of components of a mixture model on Riemannian symmetric space. The idea is to initialize the estimation process from a large number of components K_0 and introduce a prior distribution of the membership weights to express our preference for compact models. Using the Rao-Fisher information gradient to update the parameters, in each iteration the prior drives the irrelevant components to extinction. This algorithm could be applied to estimate the models on symmetric space. This algorithm is simple to use. Indeed, it is robust with respect to the choice of initial values. Moreover, it can select the number of components automatically. We illustrate this paper by some experiments.

1 Introduction

Consider a Riemannian symmertric space denoted M. For example, M may be a space of constant curvature [15], a Grassmann manifold [7], or a space of structured covariance matrices [4]. Assume that we are given $\mathcal{X} = \{x^{(1)}, \ldots, x^{(T)}\}$, a set of data points generated from an unknown probability density p_* on M.

The present work describes an original recursive algorithm, which computes an approximation of the unknown density p_*, with the assumption that it is a finite mixture density such as :

$$p_*(x|\theta) = \sum_{k=1}^{K} w_k \cdot f(x|\theta_k) \tag{1a}$$

Here $\theta = \{w_k, \theta_k; k = 1, \cdots, K\}$, where θ_k is an unknown parameter, and w_k are strictly positive membership weights. These weights satisfy the following constraint :

$$\sum_{k=1}^{K} w_k = 1 \tag{1b}$$

Moreover, each mixture component has density $f(x|\theta_k)$ chosen from a given parametric statistical model on M,

$$f(x|\theta_k) = f(x|\bar{x}_k, \sigma_k) \qquad \bar{x}_k \in M\ ,\ \sigma_k > 0 \tag{1c}$$

where $\theta_k = (\bar{x}_k, \sigma_k)$, with $\bar{x}_k$ a location parameter, and σ_k a scale parameter.

Here, since the density p_* is unknown, the order K of the mixture density p_* in (1a) is also considered to be unknown. Therefore, to compute this mixture density, one needs to solve two problems : an order selection step, to determine the order K, and a parameter estimation step, to calculate the set of weights w_k and parameters θ_k .

These two problems are studied in Sections 2 and 3, respectively. In the present work, they are solved simultaneously, and using only recursive processing. These are typical requirements when dealing with big data.

2 Order selection

As previously mentioned, the order K of the mixture density p_* is considered here to be unknown. The problem of order selection is to choose K based on the data $\mathcal{X}$. The goal here is to adapt the order K to the complexity of the available data. For example, it should not be allowed to choose a rather large value for K, if the data set $\mathcal{X}$ is essentially unimodal.

Mathematically, this is reflected by a penalty term, subtracted from the log-likelihood function $L(\theta(K)|\mathcal{X})$, in order to reject excessive values of K. The number of parameters depends on the number of components K and the notation $\theta(K)$ will be used to stress this when needed. The aim in the following will be to maximise a penalised likelihood function C given as

$$C(\theta(K)) \;=\; L(\theta|\mathcal{X}) - P(\theta(K)) \tag{2a}$$

A popular choice for the penalty term $P(\theta(K))$ arises from information criteria, such as Akaike's information criterion (AIC) [1] or the Bayesian inference criterion (BIC) [13]. However, in the present work, a different choice is made. Precisely, $P(\theta(K))$ will be derived from an non-informative Dirichlet prior on the mixture weights w_k, which reads

$$P(\theta(K)) = \log\left(\prod_{k=1}^{K} w_k^{\tilde{c}}\right) \tag{2b}$$

with $\tilde{c} = -N/2$ is equal to half the number N of real parameters θ_k involved in each mixture component. In [16], this choice of the penalty term $P(\theta(K))$ was derived as an approximation of the minimum-message-length order selection criterion.

The maximum of the penalised log-likelihood function must be a stationary point. In particular, this means its derivative with respect to the weights w_k should be 0. Introducing the Lagrange multiplier λ, to ensure the sum of the weights is always equal to 1, leads to

$$\frac{\partial}{\partial\theta_k}\left(L(\theta|\mathcal{X}) + P(\theta(K)) + \lambda(\sum_{k=1}^{K} w_k - 1)\right) = 0 \tag{2c}$$

For t data samples, we get from (2c)

$$w_k^{(t)} = \frac{1}{1 - Kc/t}\left(\frac{N_k^{(t)} - c}{t}\right) \tag{2d}$$

where $c = -\tilde{c}$ and $N_k^{(t)}$ is the expected number of samples in class k, defined in terms of the 'ownerships' $o_k^{(t)}$

$$\begin{aligned} N_k^{(t)} &= \textstyle\sum_{i=1}^{t} o_k^{(i)} \\ o_k^{(t)} &= w_k^{(t)} \frac{f(x|\theta_k)}{p(x|\theta)},\; t \in \{1, 2, \cdots, T\} \end{aligned}$$

Here, a bias from the prior is introduced through c/t, this bias decreases for larger data sets (larger t). However, if a small bias is acceptable we can keep it constant by fixing c/t to $c_T = c/T$ with a large T. This means that the bias will always be the same as if it would have been for a data set with T samples.

The interest is to look for an recursive algorithm, so every iteration the parameter updating uses only one sample. If we assume that the parameter estimate do not change much when a new sample is added, then, $o_k^{(t+1)}$ can be approximated by $o_k^{(t)}$. In this way, we can find out an update formula

$$w_k^{(t+1)} = w_k^{(t)} + \gamma^{(t+1)}\left(\frac{o_k^{(t)} - c_T}{1 - Kc_T} - w_k^{(t)}\right) \tag{2e}$$

The size of data set T should be large enough to ensure that $Kc_T < 1$. The update begins with $w_k^{(0)} = 1/K$ and discard the components whose weight becomes negative $w_k^{(t)} < 0$, until a balance could be achieved.

3 Parameter estimation

After discussing the problem of order selection in the previous section, this section will expand the problem of parameter estimation. The main parameters to estimate are the location parameter $\bar{x}_k$ defined on the Riemannian symmetric space M equipped by metric Q, and the scaling parameter $\sigma_k \in (0, +\infty)$ [10]. This kind of location-scale model has the density as follows (for each component)

$$f(x|\bar{x}_k, \sigma_k) = exp\left[\eta_k(\sigma_k)D(x, \bar{x}_k) - \psi\left(\eta_k(\sigma_k)\right)\right] \tag{3a}$$

where $\eta_k = \eta(\sigma_k)$ is a certain parameter, to be called the natural parameter, ψ is the cumulant-generating function (log-moment generating function). For example

$$\begin{aligned} \psi'(\eta) &= \mathbb{E}_\theta\left[D(x, \bar{x})\right] \\ \psi''(\eta) &= Var_\theta\left[D(x, \bar{x})\right] \end{aligned}$$

And $D : M \times M \to \mathbb{R}$ is an application which verifies the following invariance condition

$$D(g \cdot x, g \cdot \bar{x}_k) = D(x, \bar{x}_k) \tag{3b}$$

for any $g \in G$ where G is the group of isometries of M.

The estimation of the parameters can be realised by maximising the penalized log-likelihood (2a). The Fisher-information gradient will be used to do the optimisation on Riemmanian symmetric space. The Rao-Fisher information gradient is introduced by the Rao-Fisher information metric. This metric is a Riemannian metric on the parameter space $\mathcal{M} = M \times (0, +\infty)$ of the model (3a). It is defined as follows, for component k, $\theta_k = (\bar{x}_k, \sigma_k)$ and $U \in T_{\theta_k}\mathcal{M}$

$$I_{\theta_k}(U, U) = \mathbb{E}_{\theta_k}\left[\left(dL(\theta_k)U\right)^2\right] \tag{3c}$$

where $\mathbb{E}_{\theta_k}$ denotes expectation with respect to the probability density $f(x|\theta_k)$, and $dL(\theta_k)$ is the differential of the log-likelihood function, given by $L(\theta_k) = log\; f(x|\theta_k)$.

It is difficult to directly calculate the information metric by the definition (3c). So, here we need to use another metric to express the exact form of the information metric. This metric is called warped Riemannian metric as explained in [10]. A warped Riemannian metric I on $\mathcal{M}$ is given in the following way [2, 3, 9]. Let α and β be positive functions, defined on

$(0, +\infty)$. Then, for $\theta_k \in \mathcal{M}$, let the scalar product I_θ on the tangent space $T_{\theta_k}\mathcal{M}$ be defined by

$$I_{\theta_k}(U, U) = \alpha^2(\sigma_k)u_{\sigma_k}^2 + \beta^2(\sigma_k)Q_x(u, u) \tag{3d}$$

where $U = u_{\sigma_k}\partial_{\sigma_k} + u \in T_{\theta_k}\mathcal{M}$ with $u_{\sigma_k} \in \mathbb{R}$ and $u \in T_{\bar{x}_k}M$. Recall that the Q here is the metric on the space M. Theorem 1 in [10] gives the relation between information metric and warped Riemannian metric. If the model (3a) verifies the condition (3b), the Rao-Fisher information metric I of (3c) is a warped Riemannian metric given by (3c). The Proposition 1 in [10] gives the exact expression of function α and β

$$\begin{array}{rcl} \alpha^2(\eta_k) & = & \psi''(\eta_k) \\ \beta^2(\eta_k) & = & \eta_k^2 \mathbb{E}_{\theta_k}\left[Q\left(\nabla_{\bar{x}_k} D, \nabla_{\bar{x}_k} D\right) / \dim M\right] \end{array} \tag{3e}$$

where $\eta_k = \eta(\sigma_k)$. With respect to (3d) the information gradients for model (3a) could be determined [10]

$$\left\{ \begin{array}{rcl} \nabla_{\bar{x}_k}^{inf} L_k(\bar{x}_k, \eta_k) & = & \frac{1}{\beta^2(\eta_k)} \eta_k \nabla_{\bar{x}_k} Q(\bar{x}_k, x) \\ \nabla_{\eta_k}^{inf} L_k(\bar{x}_k, \eta_k) & = & \frac{1}{\psi''(\eta_k)} \left(D(\bar{x}_k, x) - \psi'(\eta_k)\right) \end{array} \right. \tag{3f}$$

With the exact expression of Rao-Fisher information gradient, an algorithm is given :

Algorithm 1 Recursive algorithm for estimating mixture models on Riemannian symmetric space

Input: data set $\{x^{(1)}, \cdots, x^{(T)}\}$, initial values, $K^{(0)}, w^{(0)}, \bar{x}^{(0)}, \sigma^{(0)}$.

Output: the estimates, $\hat{K}$, $\hat{w}$, $\hat{\bar{x}}$, $\hat{\sigma}$.

$\eta^{(0)} \leftarrow \eta(\sigma^{(0)})$.

for $t = \{1, \cdots, T\}$ **do**

  for $k = \{1, \cdots, K^{(t)}\}$ **do**

    $o_k^{(t)} \leftarrow w_k^{(t)} f(x^{(t)}|\bar{x}_k^{(t)}, \sigma_k^{(t)}) / p(x^{(t)}|\bar{x}^{(t)}, \sigma^{(t)})$.

    $w_k^{(t+1)} \leftarrow w_k^{(t)} + \gamma \left(\frac{o_k^{(t)} - c_T}{1 - K^{(t)} c_T} - w_k^{(t)} \right)$.

  end for

  if $w_k^{(t+1)} < 0, \forall k \in \{1, \cdots, K^{(t)}\}$ **then**

    $K^{(t+1)} \leftarrow K^{(t)} - 1$.

    Discard the component k.

    $w_k^{(t+1)} \leftarrow (1/K^{(t+1)}, \cdots, 1/K^{(t+1)})$.

  end if

  for $k = \{1, \cdots, K^{(t+1)}\}$ **do**

    $\bar{x}_k^{(t+1)} \leftarrow Exp_{\bar{x}_k^{(t)}}(\gamma \nabla_{\bar{x}_k^{(t)}}^{inf} L_k)$.

    $\eta_k^{(t+1)} \leftarrow Exp_{\eta_k^{(t)}}(\gamma \nabla_{\eta_k^{(t)}}^{inf} L_k)$.

  end for

end for

$\{\hat{K}, \hat{w}, \hat{\bar{x}}, \hat{\eta}\} \leftarrow \{K^{(T)}, w^{(T)}, \bar{x}^{(T)}, \eta^{(T)}\}$.

In this algorithm T denotes the number of samples, γ denotes the step size. Generaly, we use a decreasing sequence $\gamma_{t+1} = 1/(t+1)$ as the step size for stochastic gradient method, but in practice the performance of a constant step size is better. The constant coefficient c_T mentioned above equals to $\frac{N}{2T}$. In addition, Exp denotes the Riemannian exponential map. The exact definition of $Exp : TM \mapsto M$ depends on the associated manifold M. Finally, the $\nabla^{inf} L_k$ denotes the Rao-Fisher information gradient, or natural gradient.

4 Experiments

In order to verify the performance of the algorithm above, some experiments have been carried out on two models, the von Mises-Fisher distribution [6, 8] and the Riemannian Gaussian distribution [5, 11]. For lack of space, here we only present the result of Riemannian Gaussian mixture model.

For Riemannian gaussian model, M corresponds to the space of symmetric positive defined metrices with dimension $d \times d$, and the related functions in section 3 are

$$D(\bar{x}_k, x) = d^2(\bar{x}_k, x), \quad \eta(\sigma_k) = -\frac{1}{2\sigma_k^2}$$

where $d^2(\cdot, \cdot)$ denotes the Riemannian distance in M. The associated function $\psi(\eta_k)$ and $\beta(\eta_k)$ are precised in [10] and [12]. For dimension 2, the sample set $\mathcal{X}$ can be projected on the Poincaré half-plane using linear fractional transformations.

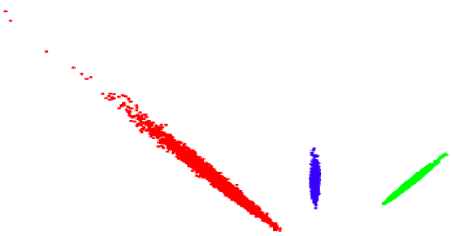

FIGURE 1 – 2×2 dimensional mixture Gaussian Riemannian distribution projected on the Poincaré half-plane

The aim is to do a simulation of estimation for this kind of mixture model. To do this simulation, the input values should be declared at first. The initial number of components $K^{(0)}$ is 10. The location parameter $\bar{x}$ is initialised by some randomly chosen data points, and the scale parameter σ is initialised by a strict random real number. The step size $\gamma = 0.01$, this constant step size brings a fluctuation at the end of the descent process. So we introduce the averaged gradient method to reduce this fluctuation [14]

$$\hat{\theta}_k^{(t+1)} = Exp_{\hat{\theta}_k^{(t)}}\left(\frac{1}{t} Log_{\hat{\theta}_k^{(t)}}(\theta_k^{(t+1)}) \right) \tag{4}$$

where $Log : \mathcal{M} \mapsto T\mathcal{M}$ is the inverse of Exp mentioned above. It provides a descent process which is smoother and easier to be observed in Figure 2.

In Figure 2, the global error is $|L(\theta^*) - L(\hat{\theta}^{(t)})|$, where θ^* denotes the true parameters, and $\hat{\theta}$ denotes the estimated parameters. The following Figure 3 presents the variation of $K^{(t)}$ with the number of samples, and the distribution of the final components $K^{(T)}$.

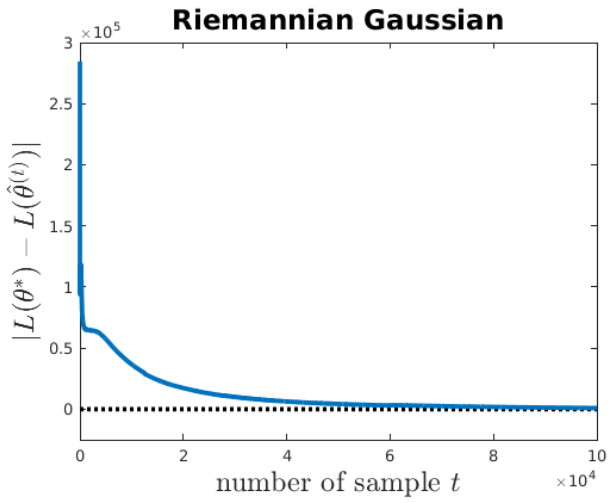

FIGURE 2 – The variation of $|L(\theta^*) - L(\theta^{(t)})|$.

From these two figures, we can observe that the errors approache to zero as the sample size increases. We have experimented 1000 simulations, 98% of the them converge to the correct number of components. This validates that the theories above are feasible in practical experiments.

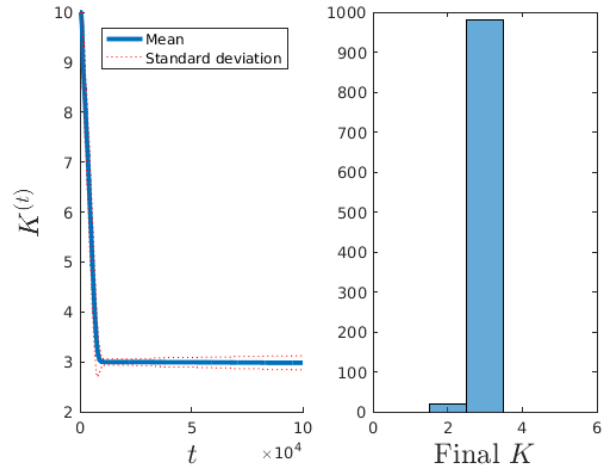

FIGURE 3 – The variation of $K^{(t)}$ and the distribution of the final $K^{(T)}$

5 Conclusion

Based on the work of [10] and [16], we have implemented an recursive algorithm. Compared with the traditional EM algorithm, this algorithm does not require complicated research on the initial values, and it can select the correct number of components automatically and estimate the parameters simultaneously. In this way, it is simple and quick to apply this algorithm to big data. In the future, the introduction of a decreasing step size may improve the accuracy of this algorithm.

Références

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6) :716–723, 1974.

[2] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[3] Richard L Bishop and Barrett O'Neill. Manifolds of negative curvature. *Transactions of the American Mathematical Society*, 145 :1–49, 1969.

[4] George Casella and Roger L Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[5] Guang Cheng and Baba C Vemuri. A novel dynamic system in the space of spd matrices with applications to appearance tracking. *SIAM journal on imaging sciences*, 6(1) :592–615, 2013.

[6] Yasuko Chikuse. *Statistics on special manifolds*, volume 174. Springer Science & Business Media, 2012.

[7] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–29. Springer, 2003.

[8] Kanti V Mardia and Peter E Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

[9] Peter Petersen, S Axler, and KA Ribet. *Riemannian geometry*, volume 171. Springer, 2006.

[10] Salem Said, Lionel Bombrun, and Yannick Berthoumieu. Warped riemannian metrics for location-scale models. In *Geometric Structures of Information*, pages 251–296. Springer, 2019.

[11] Salem Said, Lionel Bombrun, Yannick Berthoumieu, and Jonathan H Manton. Riemannian gaussian distributions on the space of symmetric positive definite matrices. *IEEE Transactions on Information Theory*, 63(4) :2153–2170, 2017.

[12] Salem Said, Hatem Hajri, Lionel Bombrun, and Baba C Vemuri. Gaussian distributions on riemannian symmetric spaces : statistical learning with structured covariance matrices. *IEEE Transactions on Information Theory*, 64(2) :752–772, 2018.

[13] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2) :461–464, 1978.

[14] Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, and Michael I Jordan. Averaging stochastic gradient descent on riemannian manifolds. *arXiv preprint arXiv :1802.09128*, 2018.

[15] Joseph Albert Wolf. *Spaces of constant curvature*, volume 372. American Mathematical Soc., 2011.

[16] Zoran Zivkovic and Ferdinand van der Heijden. Recursive unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 26(5) :651–656, 2004.