Maximum likelihood estimators on manifolds

Hatem Hajri, Salem Said, Yannick Berthoumieu

Maximum likelihood estimators on manifolds

Hatem Hajri[1], Salem Said[2], Yannick Berthoumieu[2]

[1]Institut Vedecom, 77 rue des chantiers, Versailles, [2] Laboratoire IMS (CNRS - UMR 5218), Université de Bordeaux
{hatem.hajri@vedecom.fr, {salem.said, yannick.berthoumieu }@ims-bordeaux.fr}

Abstract. Maximum likelihood estimator (MLE) is a well known estimator in statistics. The popularity of this estimator stems from its asymptotic and universal properties. While asymptotic properties of MLEs on Euclidean spaces attracted a lot of interest, their studies on manifolds are still insufficient. The present paper aims to give a unified study of the subject. Its contributions are twofold. First it proposes a framework of asymptotic results for MLEs on manifolds: consistency, asymptotic normality and asymptotic efficiency. Second, it extends popular testing problems on manifolds. Some examples are discussed.

Keywords: Maximum likelihood estimator, consistency, asymptotic normality, asymptotic efficiency of MLE, statistical tests on manifolds.

1 Introduction

Density estimation on manifolds has many applications in signal and image processing. To give some examples of situations, one can mention
Covariance matrices: In recent works [1–5], new distributions called Gaussian and Laplace distributions on manifolds of covariance matrices (positive definite, Hermitian, Toeplitz, Block Toeplitz...) are introduced. Estimation of parameters of these distributions has led to various applications (image classification, EEG data analysis, etc).
Stiefel and Grassmann manifolds: These manifolds are used in various applications such as pattern recognition [6–8] and shape analysis [9]. Among the most studied density functions on these manifolds, one finds the Langevin, Bingham and Gaussian distributions [10]. In [6–8], maximum likelihood estimations of the Langevin and Gaussian distributions are applied for tasks of activity recognition and video-based face recognition.
Lie groups: Lie groups arise in various problems of signal and image processing such as localization, tracking [11, 12] and medical image processing [13]. In [13], maximum likelihood estimation of new distributions on Lie groups, called Gaussian distributions, is performed and applications are given in medical image processing. The recent work [4] proposes new Gaussian distributions on Lie groups and a complete program, based on MLE, to learn data on Lie groups using these distributions.
The present paper is structured as follows. Section 2 focuses on consistency of MLE on general metric spaces. Section 3 discusses asymptotic

normality and asymptotic efficiency of MLE on manifolds. Finally Section 4 presents some hypothesis tests on manifolds.

2 Consistency

In this section it is shown that, under suitable conditions, MLEs on general metric spaces are consistent estimators. The result given here may not be optimal. However, in addition to its simple form, it is applicable to several examples of distributions on manifolds as discussed below.
Let (Θ, d) denote a metric space and let $\mathcal{M}$ be a measurable space with μ a positive measure on it. Consider $(\mathbb{P}_\theta)_{\theta\in\Theta}$ a family of distributions on $\mathcal{M}$ such that $\mathbb{P}_\theta(dx) = f(x,\theta)\mu(dx)$ and $f > 0$.
If $x_1, \cdots, x_n$ are independent random samples from $\mathbb{P}_{\theta_0}$, a maximum likelihood estimator is any $\hat{\theta}_n$ which solves

$$\max_\theta L_n(\theta) = L_n(\hat{\theta}_n) \text{ where } L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i, \theta)$$

The main result of this section is Theorem 1 below. The notation $\mathbb{E}_\theta[g(x)]$ stands for $\int_{\mathcal{M}} g(y) f(y,\theta)\mu(dy)$.

Theorem 1. *Assume the following assumptions hold for some* $\theta_0 \in \Theta$
(1) For all x, $f(x,\theta)$ *is continuous with respect to* θ.
(2) $\mathbb{E}_{\theta_0}[|\log f(x,\theta)|] < \infty$ *for all* $\theta, L(\theta) = \mathbb{E}_{\theta_0}[\log f(x,\theta)]$ *is continuous on* Θ *and uniquely maximized at* θ_0.
(3) For all compact K *of* Θ,

$$Q(\delta) := \mathbb{E}_{\theta_0}[\sup\{|\log f(x,\theta) - \log f(x,\theta')| : \theta, \theta' \in K, d(\theta,\theta') \le \delta\}]$$

satisfies $\lim_{\delta\to 0} Q(\delta) = 0$.
Let $x_1, \cdots, x_n, \cdots$ *be independent random samples of* $\mathbb{P}_{\theta_0}$. *For every compact* K *of* Θ, *the following convergence holds in probability*

$$\lim_{n\to\infty} \sup_{\theta\in K} |L_n(\theta) - L(\theta)| = 0$$

Assume moreover
(4) There exists a compact $K_0 \subset \Theta$ *containing* θ_0 *such that*

$$\mathbb{E}_{\theta_0}[|\sup\{\log f(x,\theta) : \theta \in K_0^c\}|] < \infty$$

and

$$\mathbb{E}_{\theta_0}[\sup\{\log f(x,\theta) : \theta \in K_0^c\}] < L(\theta_0)$$

Then, whenever $\hat{\theta}_n$ *exists and is unique for all* n, *it satisfies* $\hat{\theta}_n$ *converges to* θ_0 *in probability.*

Proof. Since L is a deterministic function, it is enough to prove, for every compact K,
(i) Convergence of finite dimensional distributions: $(L_n(\theta_1), \cdots, L_n(\theta_p))$ weakly converges to $(L(\theta_1), \cdots, L(\theta_p))$ for any $\theta_1, \cdots, \theta_p \in K$.

(ii) Tightness criterion: for all $\varepsilon > 0$,

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\big(\sup_{\theta,\theta'\in K, d(\theta,\theta')<\delta}|L_n(\theta)-L_n(\theta')|>\varepsilon\big)=0$$

Fact (i) is a consequence of the first assumption in (2) and the strong law of large numbers (SLLN). For (ii), set $F = \{(\theta,\theta')\in K^2, d(\theta,\theta')<\delta\}$ and note

$$\mathbb{P}\big(\sup_F |L_n(\theta)-L_n(\theta')|>\varepsilon\big)\le \mathbb{P}(Q_n(\delta)>\varepsilon)$$

where $Q_n(\delta)=\frac{1}{n}\sum_{i=1}^n \sup_F|\log f(x_i,\theta)-\log f(x_i,\theta')|$. By assumption (3), there exists $\delta_0>0$ such that $Q(\delta)\le Q(\delta_0)<\varepsilon$ for all $\delta\le\delta_0$. An application of the SLLN shows that, for all $\delta\le\delta_0$, $\lim_n Q_n(\delta)=Q(\delta)$ and consequently

$$\limsup_{n\to\infty}\mathbb{P}(Q_n(\delta)>\varepsilon)=\limsup_{n\to\infty}\mathbb{P}(Q_n(\delta)-Q(\delta)>\varepsilon-Q(\delta))=0$$

This proves fact (ii). Assume (4) holds. The bound

$$\mathbb{P}(\hat\theta_n\notin K_0)\le\mathbb{P}(\sup_{K_0^c}L_n(\theta)>\sup_{K_0}L_n(\theta))\le\mathbb{P}(\sup_{K_0^c}L_n(\theta)>L_n(\theta_0))$$

and the inequality $\sup_{\theta\in K_0^c}L_n(\theta)\le\frac{1}{n}\sum_{i=1}^n\sup_{\theta\in K_0^c}\log f(x_i,\theta)$ give

$$\mathbb{P}(\hat\theta_n\notin K_0)\le\mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^n\sup_{\theta\in K_0^c}\log f(x_i,\theta)>L_n(\theta_0)\bigg)$$

By the SLLN, $\limsup_n\mathbb{P}(\hat\theta_n\notin K_0)\le 1_{\{\mathbb{E}_{\theta_0}[\sup_{\theta\in K_0^c}\log f(x,\theta)]\ge L(\theta_0)\}}=0$. With $K_0(\varepsilon):=\{\theta\in K_0: d(\theta,\theta_0)\ge\varepsilon\}$, one has

$$\mathbb{P}(d(\hat\theta_n,\theta_0)\ge\varepsilon)\le\mathbb{P}(\hat\theta_n\in K_0(\varepsilon))+\mathbb{P}(\hat\theta_n\notin K_0)$$

where $\mathbb{P}(\hat\theta_n\in K_0(\varepsilon))\le\mathbb{P}(\sup_{K_0(\varepsilon)}L_n>L_n(\theta_0))$. Since L_n converges to L uniformly in probability on $K_0(\varepsilon)$, $\sup_{K_0(\varepsilon)}L_n$ converges in probability to $\sup_{K_0(\varepsilon)}L$ and so $\limsup_n\mathbb{P}(d(\hat\theta_n,\theta_0)\ge\varepsilon)=0$ using assumption (2).

2.1 Some examples

In the following some distributions which satisfy assumptions of Theorem 1 are given. More examples will be discussed in a forthcoming paper.
(i) Gaussian and Laplace distributions on $\mathcal{P}_m$. Let $\Theta=\mathcal{M}=\mathcal{P}_m$ be the Riemannian manifold of symmetric positive definite matrices of size $m\times m$ equipped with Rao-Fisher metric and its Riemannian distance d called Rao's distance. The Gaussian distribution on $\mathcal{P}_m$ as introduced in [1] has density with respect to the Riemannian volume given by $f(x,\theta)=\frac{1}{Z_m(\sigma)}\exp\big(-\frac{d^2(x,\theta)}{2\sigma^2}\big)$ where $\sigma>0$ and $Z_m(\sigma)>0$ is a normalizing factor only depending on σ.
Points (1) and (3) in Theorem 1 are easy to verify. Point (2) is proved in Proposition 9 [1]. To check (4), define $O=\{\theta: d(\theta,\theta_0)>\varepsilon\}$ and note

$$\mathbb{E}_{\theta_0}[\sup_O(-d^2(x,\theta))]\le\mathbb{E}_{\theta_0}[\sup_O(-d^2(x,\theta))1_{2d(x,\theta_0)\le\varepsilon-1}] \quad (1)$$

By the triangle inequality $-d^2(x,\theta) \leq -d(x,\theta_0)^2 + 2d(\theta,\theta_0)d(x,\theta_0) - d^2(\theta,\theta_0)$ and consequently (1) is smaller than

$$\mathbb{E}_{\theta_0}[\sup_O(2d(\theta,\theta_0)d(x,\theta_0) - d^2(\theta,\theta_0))1_{2d(x,\theta_0)\leq\varepsilon-1}]$$

But if $2d(x,\theta_0) \leq \varepsilon - 1$ and $d(\theta,\theta_0) > \varepsilon$,

$$2d(\theta,\theta_0)d(x,\theta_0) - d^2(\theta,\theta_0) < d(\theta,\theta_0)(\varepsilon - 1 - \varepsilon) < -\varepsilon$$

Finally (1) $\leq -\varepsilon$ and this gives (4) since $K_0 = O^c$ is compact.
Let $x_1, \cdots, x_n, \cdots$, ... be independent samples of $f(\cdot,\theta_0)$. The MLE based on these samples is the Riemannian mean $\hat{\theta}_n = \text{argmin}_\theta \sum_{i=1}^n d^2(x_i,\theta)$. Existence and uniqueness of $\hat{\theta}_n$ follow from [14]. Theorem 1 shows the convergence of $\hat{\theta}_n$ to θ_0. This convergence was proved in [1] using results of [15] on convergence of empirical barycenters.
(ii) Gaussian and Laplace distributions on symmetric spaces. Gaussian distributions can be defined more generally on Riemannian symmetric spaces [4]. MLEs of these distributions are consistent estimators [4]. This can be recovered by applying Theorem 1 as for $\mathcal{P}_m$. In the same way, it can be checked that Laplace distributions on $\mathcal{P}_m$ [2] and symmetric spaces satisfy assumptions of Theorem 1 and consequently their estimators are also consistent. Notice, for Laplace distributions, MLE coincides with the Riemannian median $\hat{\theta}_n = \text{argmin}_\theta \sum_{i=1}^n d(x_i,\theta)$.

3 Asymptotic normality and asymptotic efficiency of the MLE

Let Θ be a smooth manifold with dimension p equipped with an affine connection ∇ and an arbitrary distance d. Consider $\mathcal{M}$ a measurable space equipped with a positive measure μ and $(\mathbb{P}_\theta)_{\theta\in\Theta}$ a family of distributions on $\mathcal{M}$ such that $\mathbb{P}_\theta(dx) = f(x,\theta)\mu(dx)$ and $f > 0$.
Consider the following generalization of estimating functions [16].

Definition 1. *An estimating form is a function $\omega : \mathcal{M} \times \Theta \longrightarrow T^*\Theta$ such that for all $(x,\theta) \in \mathcal{M} \times \Theta$, $\omega(x,\theta) \in T^*_\theta\Theta$ and $\mathbb{E}_\theta[\omega(x,\theta)] = 0$ or equivalently $\mathbb{E}_\theta[\omega(x,\theta)X_\theta] = 0$ for all $X_\theta \in T_\theta\Theta$.*

Assume $l(x,\theta) = log(f(x,\theta))$ is smooth in θ and satisfies appropriate integrability conditions, then differentiating with respect to θ, the identity $\int_\mathcal{M} f(x,\theta)\mu(dx) = 1$, one finds $\omega(x,\theta) = dl(x,\theta)$ is an estimating form.
The main result of this section is the following

Theorem 2. *Let $\omega : \mathcal{M}\times\Theta \longrightarrow T^*\Theta$ be an estimating form. Fix $\theta_0 \in \Theta$ and let $(x_n)_{n\geq1}$ be independent samples of $\mathbb{P}_{\theta_0}$. Assume*
(i) There exist $(\hat{\theta}_N)_{N\geq1}$ such that $\sum_{n=1}^N \omega(x_n,\hat{\theta}_N) = 0$ for all N and $\hat{\theta}_N$ converges in probability to θ_0.
(ii) For all $u,v \in T_{\theta_0}\Theta$, $\mathbb{E}_{\theta_0}[|\nabla\omega(x,\theta_0)(u,v)|] < \infty$ and there exists $(e_a)_{a=1,\cdots,p}$ a basis of $T_{\theta_0}\Theta$ such that the matrix A with entries $A_{a,b} = \mathbb{E}_{\theta_0}[\nabla\omega(x,\theta_0)(e_a,e_b)]$ is invertible.

(iii) The function $R(\delta) =$

$$\mathbb{E}_{\theta_0}[\sup_{t\in[0,1],\bar{\theta}\in B(\theta_0,\delta)} |\nabla\omega(x,\gamma(t))(e_a(t),e_b(t)) - \nabla\omega(x,\theta_0)(e_a,e_b)|]$$

satisfies $\lim_{\delta\to 0} R(\delta) = 0$ *where* $(e_a, a = 1\cdots,p)$ *is a basis of* $T_{\theta_0}\Theta$ *as in (ii) and* $e_a(t), t\in[0,1]$ *is the parallel transport of* e_a *along* γ *the unique geodesic joining* θ_0 *and* $\bar{\theta}$.

Let $Log_\theta(\hat{\theta}_N) = \sum_{a=1}^p \Delta_a e_a$ *be the decomposition of* $Log_\theta(\hat{\theta}_N)$ *in the basis* $(e_a)_{a=1,\cdots,p}$. *The following convergence holds in distribution as* $N \longrightarrow \infty$

$$\sqrt{N}(\Delta_1,\cdots,\Delta_p)^T \Rightarrow \mathcal{N}(0,(A^\dagger)^{-1}\Gamma A^{-1})$$

where Γ *is the matrix with entries* $\Gamma_{a,b} = \mathbb{E}_{\theta_0}[\omega(x,\theta_0)e_a.\omega(x,\theta_0)e_b]$.

Proof. Take V a small neighborhood of θ_0 and let $\gamma : [0,1] \longrightarrow V$ be the unique geodesic contained in V such that $\gamma(0) = \theta_0$ and $\gamma(1) = \hat{\theta}_N$. Let $(e_a, a = 1\cdots,p)$ be a basis of $T_{\theta_0}\Theta$ as in (ii) and define $e_a(t), t\in[0,1]$ as the parallel transport of e_a along γ: $\frac{De_a(t)}{dt} = 0,\ t\in[0,1],\ e_a(0) = e_a$ where D is the covariant derivative along γ. Introduce

$$\omega_N(\theta) = \sum_{n=1}^N \omega(x_n,\theta) \text{ and } \ F_a(t) = \omega_N(\gamma(t))(e_a(t))$$

By Taylor formula, there exists $c_a \in [0,1]$ such that

$$F_a(1) = F_a(0) + F_a'(c_a) \tag{2}$$

Note $F_a(1) = 0, F_a(0) = \omega_N(\theta_0)(e_a)$ and $F_a'(t) = (\nabla\omega_N)(\gamma'(t),e_a(t)) = \sum_b \Delta_b(\nabla\omega_N)(e_b(t),e_a(t))$. In particular, $F_a'(0) = \sum_b \Delta_b(\nabla\omega_N)(e_b,e_a)$. Dividing (2) by $\sqrt{N}$, gives

$$-\frac{1}{\sqrt{N}}\omega_N(\theta_0)(e_a) = \frac{1}{\sqrt{N}}\sum_b \Delta_b(\nabla\omega_N)(e_b(c_a),e_a(c_a)) \tag{3}$$

Define $Y^N = \left(-\frac{1}{\sqrt{N}}\omega_N(\theta_0)(e_1),\cdots,-\frac{1}{\sqrt{N}}\omega_N(\theta_0)(e_p)\right)^\dagger$ and let A_N be the matrix with entries $A_N(a,b) = \frac{1}{N}(\nabla\omega_N)(e_a(c_a),e_b(c_a))$. Then (3) writes as $Y^N = (A_N)^\dagger(\sqrt{N}\Delta_1,\cdots,\sqrt{N}\Delta_p)^\dagger$. Since $\mathbb{E}_{\theta_0}[\omega(x,\theta_0)] = 0$, by the central limit theorem, Y^N converges in distribution to a multivariate normal distribution with mean 0 and covariance Γ. Note

$$A_{a,b}^N = \frac{1}{N}(\nabla\omega_N)(e_a,e_b) + R_{a,b}^N$$

where $R_{a,b}^N = \frac{1}{N}(\nabla\omega_N)(e_a(c_a),e_b(c_a)) - \frac{1}{N}(\nabla\omega_N)(e_a,e_b)$. By the SLLN and assumption (ii), the matrix B_N with entries $B_N(a,b) = \frac{1}{N}(\nabla\omega_N)(e_a,e_b)$ converges almost surely to the matrix A. Note $|R_{a,b}^N|$ is bounded by

$$\frac{1}{N}\sum_{n=1}^N \sup_{t\in[0,1]}\sup_{\bar{\theta}\in B(\theta_0,\delta)} |\nabla\omega(x_n,\gamma(t))(e_a(t),e_b(t)) - \nabla\omega(x_n,\theta_0)(e_a,e_b)|$$

By the SLLN, for δ small enough, the right-hand side converges to $R(\delta)$ defined in (iii). The convergence in probability of $\hat{\theta}_N$ to θ_0 and assumption (iii) show that $R_{a,b}^N \to 0$ in probability and so A_N converges in probability to A. By Slutsky lemma $((A_N^\dagger)^{-1}, Y_N)$ converges in distribution to $((A^\dagger)^{-1}, \mathcal{N}(0,\Gamma))$ and so $(A_N^\dagger)^{-1} Y_N$ converges in distribution to $(A^\dagger)^{-1}\mathcal{N}(0,\Gamma) = \mathcal{N}(0, (A^\dagger)^{-1}\Gamma A^{-1})$.

Remark 1 on $\omega = dl$. For ω an estimating form, one has $\mathbb{E}_\theta[\omega(x,\theta)] = 0$. Taking the covariant derivative, one gets $\mathbb{E}_\theta[dl(U)\omega(V)] = -\mathbb{E}_\theta[\nabla\omega(U,V)]$ for all vector fields U, V. When $\omega = dl$, this writes $\mathbb{E}_\theta[\omega(U)\omega(V)] = -\mathbb{E}_\theta[\nabla\omega(U,V)]$. In particular $\Gamma = \mathbb{E}_{\theta_0}[dl \otimes dl(e_a, e_b)] = -A$ and $A^\dagger = A = \mathbb{E}_{\theta_0}[\nabla(dl)(e_a,e_b)] = \mathbb{E}_{\theta_0}[\nabla^2 l(e_a,e_b)]$ where ∇^2 is the Hessian of l. The limit matrix is therefore equal to Fisher information matrix $\Gamma^{-1} = -A^{-1}$. This yields the following corollary.

Corollary 1. *Assume $\Theta = (M,g)$ is a Riemannian manifold and let d be the Riemannian distance on Θ. Assume $\omega = dl$ satisfies the assumptions of Theorem 2 where ∇ is the Levi-Civita connection on Θ. The following convergence holds in distribution as $N \to \infty$.*

$$N d^2(\hat{\theta}_N, \theta_0) \Rightarrow \sum_{i=1}^{p} X_i^2$$

where $X = (X_1, \cdots, X_p)^T$ is a random variable with law $\mathcal{N}(0, I^{-1})$ with $I(a,b) = \mathbb{E}_{\theta_0}[\nabla^2 l(e_a, e_b)]$.

The next proposition is concerned with asymptotic efficiency of MLE. It states that the lower asymptotic variance for estimating forms satisfying Theorem 2 is attained for $\omega_0 = dl$.
Take ω an estimating from and consider the matrices E, F, G, H with entries $E_{a,b} = \mathbb{E}_{\theta_0}[dl(\theta_0,x)e_a dl(\theta_0,x)e_b]$, $F_{a,b} = \mathbb{E}_{\theta_0}[dl(\theta_0,x)e_a \omega(\theta_0,x)e_b] = -A_{a,b}, G_{a,b} = F_{b,a}$, $H_{a,b} = \mathbb{E}_{\theta_0}[\omega(\theta_0,x)e_a\omega(\theta_0,x)e_b] = \Gamma_{a,b}$. Recall E^{-1} is the limit distribution when $\omega_0 = dl$. Note $M = \begin{pmatrix} E & F \\ G & H \end{pmatrix}$ is symmetric. When $\omega = dl$, it is furthermore positive but not definite.

Proposition 1. *If M is positive definite, then $E^{-1} < (A^\dagger)^{-1}\Gamma A^{-1}$.*

Proof. Since M is symmetric positive definite, the same also holds for its inverse. By Schur inversion lemma, $E - FH^{-1}G$ is symmetric positive definite. That is $E > FH^{-1}G$ or equivalently $E^{-1} < (A^\dagger)^{-1}\Gamma A^{-1}$.

Remark 2. As an example, it can be checked that Theorem 2 is satisfied by $\omega = dl$ of the Gaussian and Laplace distributions discussed in paragraph 2.1. For the Gaussian distribution on $\mathcal{P}_m$, this result is proved in [1]. More examples will be given in a future paper.
Remark 3 on Cramér-Rao lower bound. Assume Θ is a Riemannian manifold and $\hat{\theta}_n$ defined in Theorem 2 (i) is unbiased: $\mathbb{E}[\mathrm{Log}_{\theta_0}(\hat{\theta}_n)] = 0$. Consider $(e_1, \cdots, e_p)$ an orthonormal basis of $T_{\theta_0}\Theta$ and denote by $a = (a_1, \cdots, a_p)$ the coordinates in this basis of $\mathrm{Log}_{\theta_0}(\hat{\theta}_n)$. Smith [17] gave an intrinsic Cramér-Rao lower bound for the covariance $C(\theta_0) = \mathbb{E}[aa^T]$ as follows

$$\mathcal{C} \geq \mathcal{F}^{-1} + \text{curvature terms} \tag{4}$$

where $\mathcal{F} = (\mathcal{F}_{i,j} = \mathbb{E}[dL(\theta_0)e_i dL(\theta_0)e_j], i,j \in [1,p])$ is Fisher information matrix and $L(\theta) = \sum_{i=1}^{N} \log f(x_i, \theta)$. Define $\mathcal{L}$ the matrix with entries $\mathcal{L}_{i,j} = \mathbb{E}[dl(\theta_0)e_i dl(\theta_0)e_j]$ where $l(\theta) = \log f(x_1, \theta)$. By multiplying (4) by $\sqrt{n}$, one gets, with $y = \sqrt{n}a$,

$$\mathbb{E}[yy^T] \geq \mathcal{L}^{-1} + n \times \text{curvature terms}$$

It can be checked that as $n \to \infty$, $n \times$ curvature terms $\to 0$. Recall y converges in distribution to $\mathcal{N}(0, (A^\dagger)^{-1}\Gamma A^{-1})$. Assume it is possible to interchange limit and integral, from Theorem 2 one deduces $(A^\dagger)^{-1}\Gamma A^{-1} \geq \mathcal{L}^{-1}$ which is similar to Proposition 1.

4 Statistical tests.

Asymptotic properties of MLE have led to another fundamental subject in statistics which is testing. In the following, some popular tests on Euclidean spaces are generalized to manifolds.
Let $\Theta, \mathcal{M}$ and f be as in the beginning of the previous section.
Wald test. Given $x_1, \cdots, x_n$ independent samples of $f(., \theta)$ where θ is unknown, consider the test $H_0 : \theta = \theta_0$. Define the Wald test statistic for H_0 by

$$Q_W = n(\Delta_1, \cdots, \Delta_p) I(\theta_0) (\Delta_1, \cdots, \Delta_p)^T$$

where $I(\theta_0)$ is Fisher matrix with entries $I(\theta_0)(a,b) = -\mathbb{E}_{\theta_0}[\nabla^2 l(e_a, e_b)]$ and $\Delta_1, \cdots, \Delta_p$, $(e_a)_{a=1:p}$ are defined as in Theorem 2.
The score test. Continuing with the same notations as before, the score test is based on the statistic

$$Q_S = U(\theta_0)^T I(\theta_0) U(\theta_0)$$

where $U(\theta_0) = (U_1(\theta_0), \cdots, U_p(\theta_0))$, $(U_a(\theta_0))_{a=1:p}$ are the coordinates of $\nabla_{\theta_0} l(\theta_0, X)$ in the basis $(e_a)_{a=1:p}$ and $l(\theta, X) = \sum_{i=1}^{n} \log(f(x_i, \theta))$.

Theorem 3. *Assume $\omega = dl$ satisfies conditions of Theorem 2. Then, under $H_0 : \theta = \theta_0$, Q_W (respectively Q_S) converges in distribution to a χ^2 distribution with $p = dim(\Theta)$ degrees of freedom. In particular, Wald test (resp. the score test) rejects H_0 when Q_W (resp. Q_S) is larger than a chi-square percentile.*

Because of the lack of space, the proof of this theorem will be published in a future paper. One can also consider a generalization of Wilks test to manifolds. An extension of this test to the manifold $\mathcal{P}_m$ appeared in [1]. Future works will focus on applications of these tests to applied problems.

References

1. Said, S., Bombrun, L., Berthoumieu, Y., H. Manton, J.: Riemannian Gaussian distributions on the space of symmetric positive definite matrices. To appear in IEEE. Inf. Theory.

2. Hajri, H., Ilea, I., Said, S., Bombrun, L., Berthoumieu, Y.: Riemannian laplace distribution on the space of symmetric positive definite matrices. Entropy **18**(3) (2016)
3. Hajri, H., Bombrun, L., Said, S., Berthoumieu, Y.: A geometric learning approach on the space of complex covariance matrices. Icassp (2017)
4. Said, S., Hajri, H., Bombrun, L., Vemuri, B.C.: Gaussian distributions on Riemannian symmetric spaces: statistical learning with structured covariance matrices. available on arxiv (2016)
5. Zanini, P., Said, S., Berthoumieu, Y., Jutten, C.: Parameters estimate of Riem. gaussian distribution in the manifold of covariance matrices. IEEE Sensor Array. Rio de Janeiro. (2016)
6. Turaga, P.K., Veeraraghavan, A., Chellappa, R.: Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In: CVPR, IEEE Computer Society (2008)
7. Aggarwal, G., Roy-Chowdhury, A.K., Chellappa, R.: A system identification approach for video-based face recognition. In: ICPR (4), IEEE Computer Society (2004) 175–178
8. Turaga, P.K., Veeraraghavan, A., Srivastava, A., Chellappa, R.: Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. IEEE Trans. Pattern Anal. Mach. Intell. **33**(11) (2011) 2273–2286
9. Kendall, D.G.: Shape manifolds, procrustean metrics, and complex projective spaces. Bulletin of the London Mathematical Society (1984)
10. Chikuse, Y.: Statistics on Special Manifolds. Lecture Notes in Statistics. Vol. 174, Springer-Verlag, New York. (2003)
11. Kwon, J., Lee, H.S., Park, F.C., Lee, K.M.: A geometric particle filter for template-based visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. **36**(4) (2014) 625–643
12. Trumpf, J., Mahony, R.E., Hamel, T., Lageman, C.: Analysis of nonlinear attitude observers for time-varying reference measurements. IEEE Trans. Automat. Contr. **57**(11) (2012) 2789–2800
13. Fletcher, P.T., Joshi, S.C., Lu, C., Pizer, S.M.: Gaussian distributions on Lie groups and their application to statistical shape analysis. In: Information Processing in Medical Imaging, 18th International Conference, IPMI 2003, Ambleside, UK. (2003) 450–462
14. Afsari, B.: Riemannian L^p center of mass: existence, uniqueness and convexity. Proc. Amer. Math. Soc. **139**(2) (2011) 655–673
15. Bhattacharya, R., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds. I. Ann. Stat. **31**(1)
16. Heyde, C.C.: Quasi-likelihood and its application: a general approach to optimal parameter estimation. Springer-Verlag Inc, Berlin; New York (1997)
17. Smith, S.T.: Covariance, subspace, and intrinsic Cramér-Rao bounds. IEEE Trans. Signal Process. **53**(5) (2005) 1610–1630