Hidden Markov chains and fields with observations in Riemannian manifolds

Salem Said, Nicolas Le Bihan, Jonathan Manton

Hidden Markov chains and fields with observations in Riemannian manifolds

Salem Said * **Nicolas Le Bihan** ** **Jonathan H. Manton** ***

* *CNRS, University of Bordeaux (salem.said@u-bordeaux.fr)*
** *CNRS, Gipsa-lab (nicolas.le-bihan@gipsa-lab.grenoble-inp.fr)*
*** *The University of Melbourne (jmanton@unimelb.edu.au)*

Abstract: Hidden Markov chain, or Markov field, models, with observations in a Euclidean space, play a major role across signal and image processing. The present work provides a statistical framework which can be used to extend these models, along with related, popular algorithms (such as the Baum-Welch algorithm), to the case where the observations lie in a Riemannian manifold. It is motivated by the potential use of hidden Markov chains and fields, with observations in Riemannian manifolds, as models for complex signals and images.

Keywords: Riemannian manifold, hidden Markov model, Markov field, EM algorithm

1. INTRODUCTION

The present work is concerned with hidden Markov chain and Markov field models, with observations in Riemannian manifolds.

It introduces a general formulation of hidden Markov chain models, whose observations lie in a Riemannian manifold, and derives an expectation-maximisation algorithm, to estimate the parameters of these models — Sections 2 and 3, respectively.

It also describes a general hidden Markov field model, with observations in a Riemannian manifold, and discusses briefly the estimation of this model, using the expectation-maximisation approach — Section 5.

Hidden Markov chain and field models have been highly influential in signal and image analysis [1][2]. For example, the Baum-Welch algorithm, for estimating hidden Markov chain models, is a staple of applications such as speech recognition and protein sequencing [3] [4]. Also, hidden Markov fields are central in several approaches to image restoration and segmentation [5].

While quite extensive, existing literature is almost entirely focused on hidden Markov chain and field models with observations in a Euclidean space. The present work aims to provide a statistical framework, which may be used to extend current algorithms, dealing with observations in a Euclidean space, to observations in more general Riemannian manifolds.

For example, the expectation-maximisation algorithm, derived in Section 3, extends the popular Baum-Welch algorithm to hidden Markov chains with observations in a Riemannian manifold.

The motivation for considering hidden Markov chains and fields with observations in Riemannian manifolds comes from the important role which data in Riemannian manifolds can play in the study of complex signals and images (or even three-dimensional fields, such as turbulent flows [6]).

To understand this, consider the case of a non-stationary, multivariate signal, observed over a long time interval. This is typical of a phased-array radar, or brain-computer interface signal [7][8].

While this signal cannot be entirely described by a single tractable model, it is still possible to "zoom in" on short time windows, which admit of a complete, tractable description, very often in terms of a covariance matrix (for array radar signals, this has an additional block-Toeplitz structure [7]).

Now, the original complex signal appears as a time series of covariance matrices, or other descriptors, one for each short time window, and it seems possible to extract useful information, by studying the statistical distribution of these descriptors.

This idea can also be applied to a complex image, or three-dimensional field, by zooming in on small connected regions [9][6]. Recently, it has generated significant interest in the problem of estimating an unknown distribution, underlying a population of descriptors (*e.g.*, covariance matrices, principal subspaces, histogrammes) [10][11][12].

Mathematically, this problem was framed as the problem of estimating an unknown probability distribution on a Riemannian manifold. Indeed, in several applications, this Riemannian approach has become standard, due to its superior performance (*e.g.*, brain-computer interface [8]).

Existing work on this problem has always started from the assumption that descriptors, obtained from different time windows of a signal (or regions of an image), are sampled independently, all from the same underlying distribution. This assumption obscures the dynamics of the signal (or the spatial dependence structure within the image), potentially leading to the loss of crucial information.

The statistical framework introduced in the present work captures this information, in the form of a hidden Markov dependence structure, which controls the distribution of the descriptors.

2. HIDDEN MARKOV CHAINS

Hidden Markov chains are popular models for many kinds of signals (neuro-biological, speech, *etc.* [1][2]). Here, these models are generalised to manifold-valued signals.

In a hidden Markov chain model, one is interested in understanding some hidden, time-varying, finitely-valued, process, say $(q_t\,;t=1,2,\ldots)$ which takes values in a finite set S. When $q_t=\alpha$ for some $\alpha\in S$, one says that the process is in state α at time t.

Moreover, it is assumed that (q_t) is a time-stationary Markov chain, so there exists a so-called transition matrix, $(P_{\alpha\beta}\,;\alpha,\beta\in S)$, which gives the conditional probabilities

$$\mathbb{P}(q_{t+1}=\beta|q_t=\alpha)\;=\;P_{\alpha\beta} \tag{1}$$

In particular, if $\pi_t(\alpha)=\mathbb{P}(q_t=\alpha)$ is the distribution of q_t, then [13]

$$\pi_{t+1}(\beta)\;=\;\sum_{\alpha\in S}\pi_t(\alpha)P_{\alpha\beta} \tag{2}$$

describes the transition from time t to time $t+1$.

The states (q_t) are hidden, so they can only be observed through random outputs (y_t) which have their values in a Riemannian manifold M.

Moreover, these outputs are generated independently from each other, and each y_t depends only on q_t [1][2].

It is assumed M is a homogeneous Riemannian manifold, and y_t is distributed according to a location-scale model,

$$p(y_t|q_t=\alpha)\;=\;f(y_t|\bar{y}_\alpha\,,\sigma_\alpha) \tag{3}$$

where $p(y_t|q_t=\alpha)$ denotes the conditional density, with respect to the Riemannian volume measure of M.

The location parameters $\bar{y}_\alpha\in M$, and scale parameters $\sigma_\alpha\geq 0$ are unknown, but the function $f(y_t|\bar{y}_\alpha\,,\sigma_\alpha)$ is known, and of the form [14]

$$f(y|\bar{y},\sigma)\;=\;\exp\left[\,\eta(\sigma)D(y\,,\bar{y})-\psi(\eta(\sigma))\right] \tag{4}$$

where $D:M\times M\to\mathbb{R}$, and where $\eta(\sigma)$ is the so-called natural parameter, while $\psi(\eta)$ is a strictly convex function.

Examples of location-scale models of the form (4) include the von Mises-Fisher model, and the Riemannian Gaussian model, which are briefly recalled, in Remark 1, at the end of the present section.

From the definition of π_t and from (3), it follows that the probability density of y_t is a time-varying mixture density

$$p(y_t)\;=\;\sum_{\alpha\in S}\pi_t(\alpha)\,f(y_t|\bar{y}_\alpha\,,\sigma_\alpha) \tag{5}$$

Having access only to the observations $(y_t\,;t=1,2,\ldots)$, one hopes to recover as much information about the Markov chain (q_t) as possible: its transition matrix $(P_{\alpha\beta})$ and the parameters $(\bar{y}_\alpha\,,\sigma_\alpha)$ which govern its outputs y_t.

The problem of estimating $(P_{\alpha\beta})$ and $(\bar{y}_\alpha\,,\sigma_\alpha)$ is addressed in Section 3. A further, equally interesting problem, will not be considered, for lack of space. This is the problem of estimating the invariant distribution of the chain (q_t) [13]: if the transition matrix $(P_{\alpha\beta})$ is irreducible and aperiodic, then $\pi_t(\alpha)\to\pi(\alpha)$, the invariant distribution, as $t\to\infty$. This means that, in time, the observation y_t will tend to sample from a mixture distribution of the same form (5), but with the invariant weights $\pi(\alpha)$ instead of $\pi_t(\alpha)$.

◇ ***Remark 1 :*** location-scale models of the form (4) were introduced in [14]. For these models, M is any homogeneous Riemannian manifold [15]: a Riemannian manifold with a group of isometries G which acts transitively.

Isometry means each $g\in G$ is a mapping $g:M\to M$ which preserves the Riemannian metric of M. Moreover, transitive action means that for each $y,z\in M$ there exists $g\in G$ such that $g\cdot y=z$ (here, $g\cdot y=g(y)$).

The function $D:M\times M\to\mathbb{R}$, appearing in (4) is required to satisfy the group-invariance property

$$D(y\,,\bar{y})\;=\;D(g\cdot y\,,g\cdot\bar{y}) \tag{6}$$

for any $g\in G$ and $y,\bar{y}\in M$.

This property guarantees that $f(y|\bar{y},\sigma)$ is indeed a true probability density. In particular, the function $\psi(\eta)$ is strictly convex, because it is the cumulant generating function of the statistic $\Delta=D(y\,,\bar{y})$,

$$\mathbb{E}_{(\bar{y},\sigma)}\exp(\eta\,\Delta)\;=\;e^{\psi(\eta)}$$

where the expectation is with respect to the density $f(y|\bar{y},\sigma)$ – this identity holds for any real η as long as the expectation is finite. ■

Examples of location-scale models of the form (4) include the von Mises-Fisher model, and the Riemannian Gaussian model.

◇ ***von Mises-Fisher model*** [16]: for this model,

$$M=S^{d-1}\;\text{ and }\;G=O(d) \tag{7}$$

where $S^{d-1}\subset\mathbb{R}^d$ is the $(d-1)$-dimensional unit sphere, and $O(d)$ is the group of orthogonal transformation of $\mathbb{R}^d$.

The function $D:M\times M\to\mathbb{R}$ is given by

$$D(y\,,\bar{y})\;=\;\langle y,\bar{y}\rangle_{\mathbb{R}^d} \tag{8}$$

where $\langle\cdot,\cdot\rangle_{\mathbb{R}^d}$ is the Euclidean scalar product in $\mathbb{R}^d$, and $\psi(\eta)$ has the following expression (where $\nu=d/2$)

$$e^{\psi(\eta)}\;=\;(2\pi)^\nu\,\eta^{1-\nu}I_{\nu-1}(\eta) \tag{9}$$

where $I_{\nu-1}$ is the modified Bessel function of order $\nu-1$. Finally, the von Mises-Fisher model is obtained by setting $\eta(\sigma)=\sigma$ in (4) — then, η is called the concentration parameter. ■

◇ ***Riemannian Gaussian model*** [17][10]: for this model

$$M=\mathcal{P}_d\;\text{ and }\;G=GL(d) \tag{10}$$

where $\mathcal{P}_d$ is the space of $d\times d$ symmetric positive-definite matrices, and $GL(d)$ the group of invertible $d\times d$ matrices.

Here, G acts on M by $g\cdot y=gyg^\dagger$ where † denotes transpose. Moreover, this action preserves the Riemannian distance,

$$d^2(y,z)\;=\;\mathrm{tr}\left[\left(\log(y^{-1}z)\right)^2\right]$$

where tr denotes the trace, and log the symmetric matrix logarithm. The function $D:M\times M\to\mathbb{R}$ is then given by

$$D(y\,,\bar{y})\;=\;d^2(y,\bar{y}) \tag{11}$$

and an expression of $\psi(\eta)$, in terms of a multivariate integral, is given in [17].

The model is obtained by setting $\eta(\sigma)=-\frac{1}{2\sigma^2}$ in (4). This gives the Riemannian Gaussian density

$$f(y|\bar{y},\sigma)\;=\;Z^{-1}(\sigma)\exp\left[-\frac{d^2(y,\bar{y})}{2\sigma^2}\right] \tag{12}$$

where $Z(\sigma)=e^{\psi(\eta(\sigma))}$. ■

3. THE EM ALGORITHM

Here, an EM (expectation-maximisation) algorithm will be introduced, which addresses the estimation problem defined in Section 2.

To see why the expectation-maximisation approach is a natural choice, assume it were possible to access the so-called complete data $x_t = (y_t\,, q_t)$ at times $t = 1, \ldots, T$. The resulting log-likelihood function would be,

$$\ell(x|\theta) \;=\; \log\left[\pi_1(q_1)\prod_{t=1}^{T-1} P_{q_t q_{t+1}}\right] + \log\left[\prod_{t=1}^{T} f(y_t|\bar{y}_{q_t}\,,\sigma_{q_t})\right] \qquad (13)$$

where $\theta \;=\; (P_{\alpha\beta}\,, \bar{y}_\alpha\,, \sigma_\alpha)$. However, this could also be written [1] ,

$$\ell(x|\theta) = \sum_{\alpha,\beta\in S} N_{\alpha\beta}(q)\log\left(P_{\alpha\beta}\right) + \sum_{\alpha\in S}\sum_{t=1}^{T}\delta_\alpha(q_t)\log\, f(y_t|\bar{y}_\alpha\,,\sigma_\alpha) \quad (14)$$

where $N_{\alpha\beta}(q)$ is the number of times that $(q_t\,, q_{t+1}) = (\alpha, \beta)$, and where $\delta_\alpha(q_t) = 1$ if $q_t = \alpha$ and $= 0$ otherwise.

Estimating θ from the complete data, by maximising the log-likelihood function (14), is relatively straightforward, as explained in Remark 3 below.

Unfortunately, since the Markov chain (q_t) is hidden, one has access only to the observations $(y_t\,; t = 1, \ldots, T)$.

The log-likelihood function for these observations is much more complicated than (14),

$$\ell(y|\theta) = \log\left[\sum_{q_1\in S}\cdots\sum_{q_T\in S}\pi_1(q_1)\prod_{t=1}^{T-1}P_{q_t q_{t+1}}\prod_{t=1}^{T} f(y_t|\bar{y}_{q_t}\,,\sigma_{q_t})\right] \quad (15)$$

Instead of directly maximising (15), the expectation-maximisation approach reduces the problem of estimating θ from the observations y_t to a sequence of steps which only involve maximising a function of the form (14).

Specifically, the k-th iteration of the EM algorithm computes estimates

$$\theta^k = (\hat{P}^k_{\alpha\beta}\,, \bar{y}^k_\alpha\,, \sigma^k_\alpha)$$

which are guaranteed to increase the log-likelihood (15), in the sense that $\ell(y|\theta^k) \geq \ell(y|\theta^{k-1})$. Each iteration is made up of two steps [18][2]

• **E step** : compute the function

$$Q(\theta|\theta^{k-1}) \;=\; \mathbb{E}_{\theta^{k-1}}\left[\ell(x|\theta)|y\right] \qquad (16)$$

where $\mathbb{E}_{\theta^{k-1}}$ means the expectation is computed under the current estimate θ^{k-1} . This is called the expectation step.

• **M step** : maximise $Q(\theta|\theta^{k-1})$ to obtain the new estimate

$$\theta^k \;=\; \text{argmax}_\theta\; Q(\theta|\theta^{k-1}) \qquad (17)$$

In principle, the algorithm cycles through these two steps, until the estimates θ^k converge. In practice, this means the algorithm is designed to stop when the difference between θ^k and θ^{k-1} falls short of some pre-assigned accuracy.

There is no theoretical guarantee that the θ^k will converge to the global maximum of the log-likelihood (15). In fact, convergence typically takes place only to a local maximum, depending on initialisation.

[1] the term depending on π_1 is discarded, since it is statistically non-significant when T is large.

However, the EM algorithm remains quite advantageous, from a computational point of view, since it boils down to relatively simple operations, with little computational complexity, in comparison with alternative methods.

Indeed, replacing (14) into (16), yields $Q(\theta|\theta^{k-1}) = Q(\theta)$,

$$Q(\theta) = \sum_{\alpha,\beta\in S}\nu_{\alpha\beta}(y)\log\left(P_{\alpha\beta}\right) + \sum_{\alpha\in S}\sum_{t=1}^{T}\omega^t_\alpha(y)\log\, f(y_t|\bar{y}_\alpha\,,\sigma_\alpha) \quad (18)$$

with $\nu_{\alpha\beta}(y)$ and $\omega^t_\alpha(y)$ computed from the observations y_t ,

$$\nu_{\alpha\beta}(y) \;=\; \sum_{t=1}^{T-1}\mathbb{P}_{\theta^{k-1}}(q_t = \alpha\,, q_{t+1} = \beta|y) \qquad (19a)$$

$$\omega^t_\alpha(y) \;=\; \mathbb{P}_{\theta^{k-1}}(q_t = \alpha|y) \qquad (19b)$$

where $\mathbb{P}_{\theta^{k-1}}$ means the probability is computed under the current estimate θ^{k-1} .

Thus, the E step (16) amounts to implementing formulae (19a) and (19b). This is efficiently realised using Levinson's forward-Backward algorithm, recalled in Remark 2.

On the other hand, the M step (17) amounts to maximising the function (18), which is of the form (14), and can be maximised as explained in Remark 3.

◇ ***Remark 2 :*** in order to implement formulae (19), it is helpful to use the forward variables $\Phi_t(\alpha)$ and backward variables $\mathrm{B}_t(\alpha)$, which satisfy

$$\nu_{\alpha\beta}(y) \;=\; \sum_{t=1}^{T-1}\Phi_t(\alpha)\left(P_{\alpha\beta}f(y_{t+1}|\bar{y}_\beta\,,\sigma_\beta)\right)\mathrm{B}_{t+1}(\beta) \quad (20a)$$

$$\omega^t_\alpha(y) \;=\; \Phi_t(\alpha)\,\mathrm{B}_t(\alpha) \qquad (20b)$$

and can be computed using Levinson's forward-backward algorithm, in terms of the forward recursion

$$\Phi_{t+1}(\beta) \;=\; \Phi_t^{-1}\left[\sum_{\alpha\in S}\Phi_t(\alpha)P_{\alpha\beta}\right] f(y_{t+1}|\bar{y}_\beta\,,\sigma_\beta) \quad (21a)$$

and of the backward recursion,

$$\mathrm{B}_t(\alpha) \;=\; \Phi_{t+1}^{-1}\sum_{\beta\in S}P_{\alpha\beta}\,\mathrm{B}_{t+1}(\beta)\,f(y_{t+1}|\bar{y}_\beta\,,\sigma_\beta) \qquad (21b)$$

where Φ_t is a normalising factor,

$$\Phi_t \;=\; \sum_{\beta\in S}\left[\sum_{\alpha\in S}\Phi_t(\alpha)P_{\alpha\beta}\right] f(y_{t+1}|\bar{y}_\beta\,,\sigma_\beta) \qquad (21c)$$

and where (21a) is taken with the initial condition

$$\Phi_1(\alpha) \;=\; \frac{\pi_1(\alpha)f(y_1|\bar{y}_\alpha\,,\sigma_\alpha)}{\sum_{\alpha\in S}\pi_1(\alpha)f(y_1|\bar{y}_\alpha\,,\sigma_\alpha)} \qquad (21d)$$

and (21b) is taken with the terminal condition $\mathrm{B}_T(\beta) = 1$.

The recursions (21a) and (21b) are efficient alternatives to Baum's forward-backward algorithm, introduced in [3].

They have the same computational complexity (roughly, the order of $T|S|^2$ where $|S|$ is the number of elements in S). On the other hand, they do not suffer from the same numerical instability, which makes Baum's algorithm rather difficult to use [19]. ■

The proof of (20) and of (21) is discussed in Appendix A.

◇ ***Remark 3 :*** consider the task of maximising a function of the form (14),

$$Q(\theta) = \sum_{\alpha,\beta\in S} \nu_{\alpha\beta} \log\left(P_{\alpha\beta}\right) + \sum_{\alpha\in S}\sum_{t=1}^{T} \omega_\alpha^t \log\, f(y_t|\bar{y}_\alpha\,,\sigma_\alpha) \qquad (22)$$

with $\nu_{\alpha\beta}\,,\,\omega_\alpha^t > 0$ and $f(y|\bar{y},\sigma)$ given by (4), where the natural parameter $\eta(\sigma)$ is negative – if $\eta(\sigma)$ is positive, as in the von Mises-Fisher model, it is possible to absorb a minus sign into the function D.

Then, any global maximiser of (22), say $\hat{\theta} = (\hat{P}_{\alpha\beta}\,,\hat{y}_\alpha\,,\hat{\sigma}_\alpha)$, must satisfy

$$\hat{P}_{\alpha\beta} \,=\, \hat{P}_{\alpha\beta}(\nu)\;;\; \hat{y}_\alpha \,=\, \hat{y}_\alpha(\omega)\;;\; \eta(\hat{\sigma}_\alpha) = \hat{\eta}_\alpha(\omega)$$

where the following definitions are used

$$\hat{P}_{\alpha\beta}(\nu) \,=\, \frac{\nu_{\alpha\beta}}{\sum_{\beta\in S}\nu_{\alpha\beta}} \qquad (23a)$$

$$\hat{y}_\alpha(\omega) \,=\, \mathrm{argmin}_{y\in M} \sum_{t=1}^{T} \omega_\alpha^t\, D(y_t\,,\,y) \qquad (23b)$$

$$\hat{\eta}_\alpha(\omega) \,=\, (\psi')^{-1}\left(\frac{\sum_{t=1}^{T}\omega_\alpha^t\, D(y_t\,,\,\hat{y}_\alpha)}{\sum_{t=1}^{T}\omega_\alpha^t}\right) \qquad (23c)$$

Here, ψ' is the derivative of ψ and $(\psi')^{-1}$ its reciprocal function. In particular, $(\psi')^{-1}$ is well-defined since ψ is strictly convex. ■

The proof of Formulae (23) is discussed in Appendix B.

For the von Mises-Fisher model, the solution of the minimisation problem (23b) is given by [16]

$$\hat{y}_\alpha(\omega) \,=\, U\left(\frac{\sum_{t=1}^{T}\omega_\alpha^t\, y_t}{\sum_{t=1}^{T}\omega_\alpha^t}\right) \qquad (24a)$$

with $U : \mathbb{R}^d - \{0\} \to \mathbb{R}^d$ the mapping $U(y) = y/\|y\|_{\mathbb{R}^d}$ where $\|\cdot\|_{\mathbb{R}^d}$ denotes the Euclidean norm.

For the Riemannian Gaussian model, (23b) reads, after replacing (11),

$$\hat{y}_\alpha(\omega) \,=\, \mathrm{argmin}_{y\in M} \sum_{t=1}^{T} \omega_\alpha^t\, d^2(y_t\,,y) \qquad (24b)$$

This means that $\hat{y}_\alpha(\omega)$ is the weighted Riemannian centre of mass of the observations y_t , and can be computed using any of the standard methods for computing Riemannian centres of mass [17][10].

Formula (23c) is easily evaluated. The function $\psi'(\eta)$ being known, $\hat{\eta}_\alpha(\omega)$ is the unique solution of the equation

$$\psi'(\hat{\eta}_\alpha(\omega)) \,=\, \frac{\sum_{t=1}^{T}\omega_\alpha^t\, D(y_t\,,\,\hat{y}_\alpha)}{\sum_{t=1}^{T}\omega_\alpha^t} \qquad (25)$$

Which can be found, to any accuracy, in exponential time, using the Newton-Raphson method, since the function $\psi(\eta)$ is strictly convex.

It is also possible to find $\hat{\eta}_\alpha(\omega)$ directly, by performing a search in a table which contains the values of the function $\psi'(\eta)$. If the table is sufficiently large, this approach is successful within constant time, but has limited accuracy.

4. COMPUTER EXPERIMENT

Here, a numerical experiment is considered, illustrating the hidden Markov chain model of Section 2, and testing the validity of the EM algorithm of Section 3.

In this experiment, the hidden Markov chain (q_t) takes values in a three-element set $S = \{1, 2, 3\}$, and has initial distribution $\pi_0(\alpha) = \delta_1(\alpha)$, where $\delta_1(\alpha) = 1$ if $\alpha = 1$ and $= 0$ otherwise, and transition matrix $(P_{\alpha\beta})$,

$$(P_{\alpha\beta}) \,=\, \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{pmatrix} \qquad (26)$$

The outputs (y_t) were generated from a Riemannian Gaussian model in the Poincaré disk. In other words, each (y_t) has its values in $M = \{z \in \mathbb{C} \,:\, |z| < 1\}$, and the conditional density (3) takes on the form (12), where [20]

$$d(y,z) \,=\, \mathrm{acosh}\left[1 + \frac{2|y-z|^2}{(1-|y|^2)(1-|z|^2)}\right] \qquad (27a)$$

$$Z(\sigma) \,=\, (2\pi)^{3/2}\,\sigma\, e^{\frac{\sigma^2}{2}}\,\mathrm{erf}\left(\frac{\sigma}{\sqrt{2}}\right) \qquad (27b)$$

erf being the error function [10].

The Poincaré disk is considered here, rather than a space $\mathcal{P}_d$ of positive-definite matrices, as in Section 2, since it is much easier to visualise (indeed, it is just the inside of a unit disk in the plane). On the other hand, the Poincaré disk, with Riemannian distance (27a) is isometric to the space of 2×2 symmetric positive-definite matrices, with unit determinant [20].

Some background on the Riemannian geometry of the Poincaré disk is recalled in Remark 4.

The following sets of location parameters $\bar{y}_\alpha$ and scale parameters σ_α were used

$$\bar{y}_1 \,=\, 0\;;\; \bar{y}_2 \,=\, 0.82\,\mathrm{i} + 0.29\;;\; \bar{y}_3 \,=\, 0.82\,\mathrm{i} - 0.29 \qquad (28a)$$

$$\sigma_1 \,=\, 0.1\;;\; \sigma_2 \,=\, 0.4\;;\; \sigma_3 \,=\, 0.4 \qquad (28b)$$

where $\mathrm{i} = \sqrt{-1}$.

With the hidden Markov model defined in this way, the observations $(y_t\,;t = 1,\ldots,T)$ where $T = 10000$ appeared as a "face" pattern, within the Poincaré disk, shown in Figure 1.

If the scale parameters σ_α are all multiplied by 5, the face becomes more blurred, since each conditional density (3) has its dispersion multiplied by 5. This can be seen in Figure 2.

In both figures 1 and 2, the "eyes" (centred at $\bar{y}_2$ and $\bar{y}_3$) appear smaller than the "mouth", although are they have scale parameters σ_2 and σ_3 four times larger (see (28b) !).

This apparent contradiction is due to the fact that the Riemannian metric of the Poincaré disk is stretched near the boundary of the disk — see Formula (29), Remark 4. Thus, while they appear smaller to us, objects near the boundary are, in fact, bigger when measured by this Riemannian metric.

Consider now the application of the EM algorithm of Section 3, to the estimation of the transition matrix (26) and location and scale parameters (28).

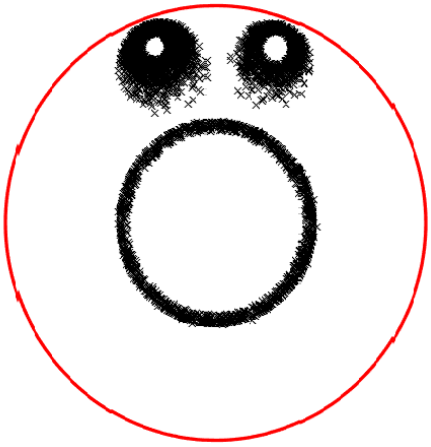

Fig. 1. 10^4 observations ; location-scale parameters (28)

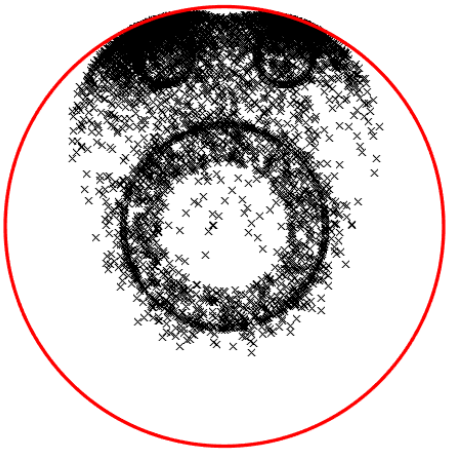

Fig. 2. 10^4 observations ; scale parameters $\times$ 5

$\diamond$ ***Remark 4:*** the Poincaré disk is the set of complex numbers z with $|z| < 1$, equipped with a conformal Riemannian metric, which turns it into a space of constant negative curvature [20].

This Riemannian metric measures the length of a plane vector w (identified with a complex number), attached at the point z, using the Riemannian norm

$$\|w\|_z^2 \;=\; \left(\frac{4}{\kappa}\right) \frac{|w|^2}{(1-|z|^2)^2} \qquad (29)$$

where $\kappa > 0$ is a scaling factor, and $|w|^2 = ww^*$ is the squared modulus, or squared Euclidean norm (the star denotes the conjugate).

It is called a conformal metric, since the Riemannian norm $\|w\|_z$ is proportional to the Euclidean norm $|w|$. The sectional curvature associated to this metric is constant, and equal to $-\kappa$.

In (29), the conformal factor $(1-|z|^2)^{-2}$ goes to infinity as $|z|$ approaches 1. Accordingly, the Riemannian norm $\|w\|_z$ of a vector w, with fixed Euclidean norm $|w|$, becomes arbitrarily large, when z approaches the boundary of the disc.

In other words, the Riemannian metric of the Poincaré disk is stretched near the boundary. ■

This algorithm was run $N_{\rm mc} = 20$ times, each time for $N_{\rm em} = 1000$ iterations. An individual run took about one hour, on a standard laptop.

All of these runs were carried out with the same initial guess θ^1, but the observations $(y_t \,; t = 1, \ldots, T)$ were generated anew for each run, leading to a different (in fact, random) value of the final estimate

$$\theta^{N_{\rm em}} \;=\; (\hat{P}_{\alpha\beta}^{N_{\rm em}}, \bar{y}_\alpha^{N_{\rm em}}, \sigma_\alpha^{N_{\rm em}})$$

Then, the mean value of $(\hat{P}_{\alpha\beta}^{N_{\rm em}})$ was found to be

$$\mathbb{E}_{\rm mc}(\hat{P}_{\alpha\beta}^{N_{\rm em}}) \;=\; \begin{pmatrix} 0.5 & 0.2 & 0.3 \\ 0.2 & 0.8 & 0 \\ 0.1 & 0 & 0.9 \end{pmatrix}$$

while the mean values of $\bar{y}_\alpha^{N_{\rm em}}$ and $\sigma_\alpha^{N_{\rm em}}$ turned out to be

$$\begin{aligned} \mathbb{E}_{\rm mc}(\bar{y}_1^{N_{\rm em}}) \;&=\; -0.02\,{\rm i} + 0.41 \;\; ; \;\; \mathbb{E}_{\rm mc}(\sigma_1^{N_{\rm em}}) \;=\; 0.5 \\ \mathbb{E}_{\rm mc}(\bar{y}_2^{N_{\rm em}}) \;&=\; \;\; 0.83\,{\rm i} + 0.30 \;\; ; \;\; \mathbb{E}_{\rm mc}(\sigma_2^{N_{\rm em}}) \;=\; 0.3 \\ \mathbb{E}_{\rm mc}(\bar{y}_3^{N_{\rm em}}) \;&=\; \;\; 0.83\,{\rm i} - 0.29 \;\; ; \;\; \mathbb{E}_{\rm mc}(\sigma_3^{N_{\rm em}}) \;=\; 0.3 \end{aligned}$$

Here, $\mathbb{E}_{\rm mc}$ is the "Monte Carlo expectation", equal to the empirical average, taken over $N_{\rm mc}$ runs.

The dispersion of $(\hat{P}_{\alpha\beta}^{N_{\rm em}})$, about its mean value, was measured by the maximum empirical variance of the entries $\hat{P}_{\alpha\beta}^{N_{\rm em}}$. This was found to be

$$\max_{\alpha\beta} \mathbb{V}_{\rm mc}(\hat{P}_{\alpha\beta}^{N_{\rm em}}) \;=\; 0.06$$

The dispersions of $\bar{y}_\alpha^{N_{\rm em}}$ and $\sigma_\alpha^{N_{\rm em}}$ were computed similarly

$$\max_{\alpha} \mathbb{V}_{\rm mc}(\bar{y}_\alpha^{N_{\rm em}}) \;=\; 0.05 \;\; ; \;\; \max_{\alpha} \mathbb{V}_{\rm mc}(\sigma_\alpha^{N_{\rm em}}) \;=\; 0.1$$

where $\mathbb{V}_{\rm mc}$ is the "Monte Carlo variance", equal to the empirical variance, taken over $N_{\rm mc}$ runs.

These numerical results show that the value of the final estimate $\theta^{N_{\rm em}}$ is rather stable, and is not affected very strongly by the randomness of the observations. Indeed, the dispersions of $\hat{P}_{\alpha\beta}^{N_{\rm em}}$ or of $\bar{y}_\alpha^{N_{\rm em}}$ and $\sigma_\alpha^{N_{\rm em}}$ appear small in comparison to their mean values.

Building on this remark, it is possible to infer that the discrepancy, between these mean values and the true values of the transition probabilities $P_{\alpha\beta}$, and location and scale parameters $\bar{y}_\alpha$ and σ_α, is due to the insufficient number of observations, $T = 10000$, rather than to a slow convergence of the EM algorithm.

This is supported by the fact that each run of the EM algorithm was observed to converge to the final estimate $\theta^{N_{\rm em}}$ well before $N_{\rm em} = 1000$ iterations, producing nearly constant estimates $\theta^k \approx \theta^{N_{\rm em}}$ not later than $k = 800$.

When the number of observations was increased to $T = 50000$, an individual run, of $N_{\rm em} = 1000$ iterations of the algorithm, took much longer, over five hours.

For this reason, only one run of the algorithm was carried out. This was found to produce final estimates significantly closer to the true values of the transition matrix and location and scale parameters.

However, for lack of time, a more detailed statistical analysis was not possible, in this case.

5. HIDDEN MARKOV FIELDS

Similar to hidden Markov chain models, hidden Markov field models involve a hidden stochastic process [2] (q_z), which takes its values in a finite set S, observed through random outputs (y_z), which belong to a homogeneous Riemannian manifold M.

However, in the case of a hidden Markov field, the index z is not a time variable (running through $t = 1, 2, \ldots$), but a vertex of a finite unidirected graph $\mathcal{Z}$. Then, the stochastic process $(q_z \,; z \in \mathcal{Z})$ is called a Markov field if the following two properties are satisfied [22]

$$\mathbb{P}(q_z = \alpha | q_w \,; w \neq z) \;=\; \mathbb{P}(q_z = \alpha | q_w \,; w \sim z) \tag{30a}$$

where $w \sim z$ means w and z are adjacent, and

$$\mathbb{P}\left(\cap_{z \in \mathcal{Z}} \{q_z = \varsigma(z)\}\right) \;> 0 \tag{30b}$$

for any function $\varsigma : \mathcal{Z} \to S$. Such a function is called a configuration of the field.

Property (30a) says that the state q_z at a vertex z depends on states at the other vertices q_w only through those w which are adjacent to z.

On the other hand, Property (30b) is required for the Hammersley-Clifford theorem [22]. This theorem implies that the probability distribution of a Markov field is a Gibbs distribution.

For simplicity, it is here assumed that the graph $\mathcal{Z}$ is a square grid, where each vertex is adjacent to its immediate neighbors (this is a simple model of an image). In this case, the field (q_z) is said to have a Gibbs distribution, if the probability mass function of the states q_z takes on the form

$$p(q_z \,; z \in \mathcal{Z}) \;\propto\; \prod_{z \in \mathcal{Z}} \exp\left[V(q_z) + \frac{1}{2} \sum_{w \sim z} J(q_z \,, q_w) \right] \tag{31}$$

Here, $\propto$ indicates a missing normalising factor $W_{(V,J)}$. Computing $W_{(V,J)}$ requires summing over all possible configurations of the field q — often, an impracticable task.

The observed outputs (y_z) depend on the hidden states (q_z) exactly as in Section 2. That is,

$$p(y_z | q_z = \alpha) \;=\; f(y_z | \bar{y}_\alpha \,, \sigma_\alpha) \tag{32}$$

where the function $f(y_z | \bar{y}_\alpha \,, \sigma_\alpha)$ is of the form (4). Then, having access only to these outputs (y_z), one hopes to recover the dependence structure of the Hidden field, given by the functions $V : S \to \mathbb{R}$ and $J : S \times S \to \mathbb{R}$, as well as the parameters $(\bar{y}_\alpha \,, \sigma_\alpha)$ which govern the (y_z).

This problem generalises the hidden Markov chain model problem of Section 2, which was addressed in Section 3. It is a problem of parameter estimation, for the unknown parameter $\theta = (V_\alpha \,, J_{\alpha\beta} \,, \bar{y}_\alpha \,, \sigma_\alpha)$, where $V_\alpha = V(\alpha)$ and $J_{\alpha\beta} = J(\alpha, \beta)$ for $\alpha, \beta \in S$.

In the special case where the outputs y_z belong to a Euclidean space $M = \mathbb{R}^d$, several EM algorithms were introduced in the image analysis literature, in order to address this problem [23][24].

[2] a stochastic process is any family of jointly defined random variables [21].

The extension of these algorithms, to observed outputs y_z in any homogeneous Riemannian manifold M, is here discussed briefly.

Each iteration of the EM algorithm includes an E step and an M step. Here, these take on the following form, where the notation is the same as in (16) and (17) of Section 3 — unfortunately, due to lack of space, the following discussion does not include any proofs.

• **E step** : compute the function

$$Q(\theta | \theta^{k-1}) \;=\; Q(\bar{y}_\alpha \,, \sigma_\alpha) \;+\; Q(V, J) \tag{33}$$

in terms of the following formulae

$$Q(\bar{y}_\alpha \,, \sigma_\alpha) \;=\; \sum_{\alpha \in S} \sum_{z \in \mathcal{Z}} \omega_\alpha^z(y) \log\, f(y_z | \bar{y}_\alpha \,, \sigma_\alpha) \tag{34a}$$

$$\begin{aligned} Q(V, J) \;=\; & -\log W(V, J) \,+ \\ & \sum_{\alpha \in S} \omega_\alpha(y) V_\alpha + \frac{1}{2} \sum_{\alpha, \beta \in S} \nu_{\alpha\beta}(y) J_{\alpha\beta} \end{aligned} \tag{34b}$$

which involve conditional probabilities and expectations,

$$\omega_\alpha^z(y) \;=\; \mathbb{P}_{\theta^{k-1}}(q_z = \alpha | y) \;;\; \omega_\alpha(y) \;=\; \sum_{z \in \mathcal{Z}} \omega_\alpha^z(y) \tag{35a}$$

$$\nu_{\alpha\beta}(y) \;=\; \sum_{w \sim z} \mathbb{P}_{\theta^{k-1}}(q_z = \alpha \,, q_w = \beta | y) \tag{35b}$$

• **M step** : maximise $Q(\theta | \theta^{k-1})$ to obtain new estimates, $\theta^k = (\hat{V}_\alpha \,, \hat{J}_{\alpha\beta} \,, \hat{y}_\alpha \,, \hat{\sigma}_\alpha)$. These are given by,

$$(\hat{V}, \hat{J}) \;=\; \mathrm{argmax}_{(V,J)} \, \langle \omega, V \rangle + \langle \nu, J \rangle - \Psi(V, J) \tag{36a}$$

and, in the notation of Remark 3,

$$\hat{y}_\alpha \;=\; \mathrm{argmin}_{\hat{y} \in M} \sum_{z \in \mathcal{Z}} \omega_\alpha^z(y) \, D(y_z \,, \hat{y}) \tag{36b}$$

$$\eta(\hat{\sigma}_\alpha) \;=\; (\psi')^{-1} \left(\frac{\sum_{z \in \mathcal{Z}} \omega_\alpha^z(y) \, D(y_z \,, \hat{y}_\alpha)}{\sum_{z \in \mathcal{Z}} \omega_\alpha^z(y)} \right) \tag{36c}$$

Here, $\Psi(V, J) = \log\, W(V, J)$, and

$$\langle \omega, V \rangle = \sum_{\alpha \in S} \omega_\alpha(y) V_\alpha \;\;;\;\; \langle \nu, J \rangle = \frac{1}{2} \sum_{\alpha, \beta \in S} \nu_{\alpha\beta}(y) J_{\alpha\beta}$$

Moreover, the maximum in (36a) is unique, since $\Psi(V, J)$ is a strictly convex function of $V = (V_\alpha)$ and $J = (J_{\alpha\beta})$, (when these are considered as $V \in \mathbb{R}^{|S|}$ and $J \in \mathbb{R}^{|S| \times |S|}$).

◇ ***Remark 5*** *:* in its form specified by (34) and (36), the EM algorithm is not practically applicable. Indeed, both the normalising constant $W(V, J)$, and the conditional probabilities and expectations (35), are impossible to evaluate, outside of trivial situations.

To circumvent this (fundamental) difficulty, various approximate methods, for evaluating $W(V, J)$ and (35), have been proposed, such as the ones based on mean-field approximations [23][24]. ■

It will be the aim of the authors' future work to adapt and apply these methods, to the general case where the observed outputs y_z belong to a homogeneous Riemannian manifold.

Appendix A. FORWARD-BACKWARD VARIABLES

Define the forward variables $\Phi_t(\alpha)$ and backward variables $\mathrm{B}_t(\alpha)$ by [19] — for convenience, write $\mathbb{P}$ instead of $\mathbb{P}_{\theta^{k-1}}$

$$\Phi_t(\alpha) \;=\; \mathbb{P}\,(q_t = \alpha | y_t\,, \ldots, y_1) \tag{A.1}$$

$$\mathrm{B}_t(\alpha) \;=\; \frac{\mathbb{P}\,(y_T\;\ldots, y_{t+1} | q_t = \beta)}{p(y_T\;\ldots, y_{t+1} | y_t\,, \ldots, y_1)} \tag{A.2}$$

To see that these satisfy (20), the key is to make use of the fact that $x_t = (y_t\,, q_t)$ is a Markov chain, with transition probabilities [1][2]

$$p(y_{t+1}\,, q_{t+1} = \beta | y_t\,, q_t = \alpha) \;=\; P_{\alpha\beta}\, f(y_{t+1} | \bar{y}_\beta\,, \sigma_\beta) \tag{A.3}$$

For example, (20a) can be obtained by using Bayes rule to express each term in the sum (19a)

$$\mathbb{P}(q_t = \alpha\,, q_{t+1} = \beta | y) \;=\; \frac{p(y_T\,, \ldots, y_1\,; q_t = \alpha\,, q_{t+1} = \beta)}{p(y_T\,, \ldots, y_1)} \tag{A.4}$$

where the numerator is the joint probability density of $y_T\,, \ldots, y_1$ on the event $(q_t\,, q_{t+1}) = (\alpha, \beta)$ — this is a density with respect to the Riemannian volume measure of the product manifold $M^T = M \times \ldots \times M$.

This joint density can be written as a product,

$$\begin{aligned} p(y_T\,, \ldots, y_{t+2} |\, y_{t+1}\,, q_{t+1} = \beta\,, q_t = \alpha\,, y_t\,, \ldots, y_1) \;\times \\ p(y_{t+1}\,, q_{t+1} = \beta \,|\, q_t = \alpha\,, y_t\,, \ldots, y_1) \;\times \\ \mathbb{P}(q_t = \alpha |\, y_t\,, \ldots, y_1) \;\times \\ p(y_t\,, \ldots, y_1) \end{aligned}$$

by repeatedly applying (again!) Bayes rule. However, by the Markov property of $x_t = (y_t\,, q_t)$, this product becomes

$$\begin{aligned} p(y_T\,, \ldots, y_{t+2} |\, y_{t+1}\,, q_{t+1} = \beta) \;\times \\ p(y_{t+1}\,, q_{t+1} = \beta \,|\, q_t = \alpha) \;\times \\ \mathbb{P}(q_t = \alpha |\, y_t\,, \ldots, y_1) \;\times \\ p(y_t\,, \ldots, y_1) \end{aligned}$$

but, using (A.1), (A.2) and (A.3), this is the same as

$$\begin{aligned} \mathrm{B}_{t+1}(\beta) \;\times \\ P_{\alpha\beta} f(y_{t+1} | \bar{y}_\beta\,, \sigma_\beta) \;\times \\ \Phi_t(\alpha) \;\times \\ p(y_t\,, \ldots, y_1) \end{aligned}$$

Therefore, replacing back into (A.4), it follows each term in the sum (19a) is given by

$$\Phi_t(\alpha)\,(P_{\alpha\beta} f(y_{t+1} | \bar{y}_\beta\,, \sigma_\beta))\,\mathrm{B}_{t+1}(\beta)$$

Thus, summing over $t = 1\,, \ldots, T-1$ yields (20a).

An analogous, and simpler, reasoning can be used to obtain (20b).

In addition to (20), it should be proved that the forward and backward variables are given by (21).

To obtain (21a), note first that the initial condition (21d) follows directly from (A.1). In addition, also from (A.1)

$$\begin{aligned} \Phi_{t+1}(\beta) \;&\propto\; \mathbb{P}\,(q_{t+1} = \beta | y_{t+1}\,, \ldots, y_1) \\ &\propto\; \mathbb{P}\,(y_{t+1}\,, q_{t+1} = \beta | y_t\,, \ldots, y_1) \end{aligned}$$

where $\propto$ indicates a missing normalising factor. Thus, using the Markov property of $x_t = (y_t\,, q_t)$, it is seen that, up to normalisation, $\Phi_{t+1}(\beta)$ is equal to

$$\begin{aligned} &\sum_{\alpha \in S} p(y_{t+1}\,; q_{t+1} = \beta | q_t = \alpha) \times \mathbb{P}(q_t = \alpha | y_t\,, \ldots, y_1) \\ &= \sum_{\alpha \in S} P_{\alpha\beta}\, f(y_{t+1} | \bar{y}_\beta\,, \sigma_\beta) \times \Phi_t(\alpha) \end{aligned}$$

This immediately gives (21a), after noting that the missing normalising factor can be found by summing over $\beta \in S$, as in (21c).

The proof of (21b) is similar to that of (21a), and is here omitted to avoid repetition.

Appendix B. PROOF OF FORMULAE (23)

Recall the function $Q(\theta)$ of (22),

$$Q(\theta) = \sum_{\alpha, \beta \in S} \nu_{\alpha\beta} \log\,(P_{\alpha\beta}) \;+\; \sum_{\alpha \in S} \sum_{t=1}^{T} \omega_\alpha^t \log\, f(y_t | \bar{y}_\alpha\,, \sigma_\alpha)$$

$\diamond$ ***Proof of (23a)*** *:* only the first term in the expression of $Q(\theta)$ depends on $(P_{\alpha\beta})$. Therefore, this term can be maximised separately.

Denote this first term by $Q(P)$. Then, maximising $Q(P)$ is equivalent to maximising

$$\tilde{Q}(P) \;=\; \sum_{\alpha, \beta \in S} \hat{P}_{\alpha\beta}(\nu) \log\,(P_{\alpha\beta})$$

where $\hat{P}_{\alpha\beta}(\nu)$ is given by (23a).

Let $\hat{P}_{\alpha\beta} = \hat{P}_{\alpha\beta}(\nu)$. For each $\alpha \in S$, let $\hat{P}_\alpha = (\hat{P}_{\alpha\beta}\,; \beta \in S)$ and $P_\alpha = (P_{\alpha\beta}\,; \beta \in S)$. Both $\hat{P}_\alpha$ and P_α are probability distributions on S. If $\mathrm{KL}(\hat{P}_\alpha \| P_\alpha)$ is the Kullback-Leibler divergence between $\hat{P}_\alpha$ and P_α, then [25]

$$\tilde{Q}(P) \;=\; \sum_{\alpha \in S} \Big\{ H(\hat{P}_\alpha) \;-\; \mathrm{KL}(\hat{P}_\alpha \| P_\alpha) \Big\}$$

where $H(\hat{P}_\alpha) = \sum_{\beta \in S} \hat{P}_{\alpha\beta} \log(\hat{P}_{\alpha\beta})$ is the entropy of $\hat{P}_\alpha$ (which does not depend on $(P_{\alpha\beta})$).

Therefore, $\tilde{Q}(\theta)$ reaches its maximum when each each $\mathrm{KL}(\hat{P}_\alpha \| P_\alpha)$ reaches its minimum. But this happens when $\mathrm{KL}(\hat{P}_\alpha \| P_\alpha) = 0$ at $P_{\alpha\beta} = \hat{P}_{\alpha\beta}$, as required. ■

$\diamond$ ***Proof of (23b) and (23c)*** *:* denote the second term in the expression of $Q(\theta)$ by $Q(\bar{y}, \sigma)$ and consider the task of maximising $Q(\bar{y}, \sigma)$.

This can be carried out by maximising each one of the functions

$$Q(\bar{y}_\alpha\,, \sigma_\alpha) \;=\; \sum_{t=1}^{T} \omega_\alpha^t \log\, f(y_t | \bar{y}_\alpha\,, \sigma_\alpha)$$

with respect to $\bar{y}_\alpha$ and σ_α. From (4),

$$Q(\bar{y}_\alpha\,, \sigma_\alpha) \;=\; \sum_{t=1}^{T} \omega_\alpha^t \,\{\, \eta(\sigma_\alpha) D(y_t\,, \bar{y}_\alpha) - \psi(\eta(\sigma_\alpha))\}$$

By first maximising over $\bar{y}_\alpha$ and recalling that $\eta(\sigma_\alpha)$ is negative, it is seen than any maximiser $\hat{y}_\alpha$ must be equal to $\hat{y}_\alpha(\omega)$ given by (23b).

Then, it remains to maximise (over σ_α) the function

$$Q(\sigma_\alpha) \,=\, \sum_{t=1}^{T} \omega_\alpha^t \,\{\, \eta(\sigma_\alpha) D(y_t \,, \hat{y}_\alpha) - \psi(\eta(\sigma_\alpha))\}$$

This is a strictly concave function of $\eta(\sigma_\alpha)$, since $\psi(\eta)$ is a strictly convex function. Thus, it is possible to maximise by differentiating with respect to $\eta(\sigma_\alpha)$ and setting the derivative to zero. This yields the equation

$$\sum_{t=1}^{T} \omega_\alpha^t \,\{\, D(y_t \,, \hat{y}_\alpha) - \psi'(\eta(\hat{\sigma}_\alpha))\} \,=\, 0$$

for the maximiser $\hat{\sigma}_\alpha$. This equation is equivalent to (after dividing both sides by $\sum_{t=1}^{T} \omega_\alpha^t$)

$$\psi'(\eta(\hat{\sigma}_\alpha)) \,=\, \frac{\sum_{t=1}^{T} \omega_\alpha^t \, D(y_t \,, \, \hat{y}_\alpha)}{\sum_{t=1}^{T} \omega_\alpha^t}$$

which is the same as equation (25), whose unique solution is $\eta(\hat{\sigma}_\alpha) = \hat{\eta}_\alpha(\omega)$ given by (23c). ■

REFERENCES

[1] R. J. Elliot, L. Aggoun, and J. B. Moore. *Hidden Markov Models : Estimation and Control.* Springer Science+Business Media, 1995.

[2] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models.* Springer Science+Business Media, 2005.

[3] L. R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. (In Readings in Speech Recognition).* Morgan Kaufmann Publishers, Inc, 1990.

[4] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence analysis.* Cambridge University Press, 1998.

[5] S. Z, Li. *Markov Random Field Modeling in Image Analysis.* Springer Publishing Company, 2009.

[6] A. Zare, M. Jovanovic, and T. Georgiou. *Colour of Turbulence.* Journal of Fluid Mechanics, 812:630–680, 2017.

[7] B. Jeuris, and R. Vandebril. *The Kähler mean of block Toeplitz matrices with Toeplitz structured blocks.* SIAM Journal on Matrix Analysis and Applications, 37:1151–1175, 2016.

[8] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten. *Multiclass brain-computer interface classification by Riemannian geometry.* IEEE Transactions on Biomedical Engineering, 59:920–928, 2012.

[9] O. Tuzel, F. Porikli, and P. Meer. *Pedestrian detection via classification on Riemannian manifolds.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 30:1713–1727, 2008.

[10] S. Said, H. Hajri, L. Bombrun, and B. C. Vemuri. *Gaussian distributions on Riemannian symmetric spaces : statistical learning with structured covariance matrices.* IEEE Transactions on Information Theory, 64:752–772, 2018.

[11] E. Chevallier, T. Forget, F. Barbaresco, and J. Angulo. *Kernel density estimation on the Siegel space with application to Radar processing.* Entropy, 18, 2016.

[12] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. *Clustering on the unit hypersphere using von Mises-Fisher distributions.* Journal of Machine Learning Research, 6: 1345–1382, 2005.

[13] J. R. Norris. *Markov Chains.* Cambridge University Press, 1997.

[14] S. Said, L. Bombrun, and Y. Berthoumieu. *Warped Riemannian metrics for location-scale models. (In Geometric Structures of Information).* Springer, Cham, 2019.

[15] A. L. Besse. *Einstein manifolds.* Springer-Verlag, 2002.

[16] K. V. Mardia, and P. E. Jupp. *Directional statistics.* John Wiley & Sons, 2000.

[17] S. Said, L. Bombrun, Y. Berthoumieu, and J. H. Manton. *Riemannian Gaussian distributions in the space of symmetric positive definite matrices.* IEEE Transactions on Information Theory, 63:2153–2170, 2017.

[18] A. P. Dempster, N. M. Laird, and D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society B, 39:1–38, 1977.

[19] P. Devijver. *Baum's forward-backward algorithm revisited.* Pattern Recognition Letters, 3:369–373, 1985.

[20] E. B. Vinberg. *Geometry II : Spaces of constant curvature.* Encyclopedia of Mathematical Sciences, Vol. 29. Springer Verlag, 1993.

[21] O. Kallenberg. *Foundations of modern probability.* Springer-Verlag, 2002.

[22] G. Grimmett. *Probability on graphs : stochastic processes on graphs and lattices.* Cambridge University Press, 2010.

[23] B. Celeux., F. Forbes, and N. Peyrard. *EM procedures using mean-field-like approximations for Markov-model-based image segmentation.* Research Report 4105. INRIA, 2001.

[24] B. Celeux., F. Forbes, and N. Peyrard. *EM-based image segmentation using Potts models with external field.* Research Report 4456. INRIA, 2002.

[25] T. M. Cover, and J. A. Thomas. *Elements of Information Theory.* Wiley-Blackwell, 2006.