# Learning Riemannian Metrics for Classification of Dynamical Models

Fabio Cuzzolin, Stefano Soatto

# Learning Riemannian Metrics for Classification of Dynamical Models

Fabio Cuzzolin and Stefano Soatto

December 17, 2005

**Abstract**

Consider the problem of classifying motions, encoded as dynamical models of a certain class. Standard nearest neighbor classification then reduces to find a suitable distance function in the space of the models. In this paper we present a supervised differential-geometric method to learn a Riemannian metric for a given class of dynamical models in order to improve classification performances. Given a training set of models the optimal metric is selected among a family of pullback metrics induced by the Fisher information tensor through a parameterized diffeomorphism. Experimental results concerning action and identity recognition based on simple scalar features are shown, proving how learning a metric actually improves classification rates when compared with Fisher geodesic distance and other classical distance functions.

## 1 Introduction

Human motion recognition is one of the most popular fields in computer vision, due to both its applicative potential and its richness in terms of the technical issues involved. Consider then the problem of classifying a number of movements, represented as sequences of image features: Representing those sequences in a compact way as simple dynamical models has proved to be effective in problems like dynamic textures or gait identification.

The motion classification problem then reduces to finding the appropriate distance function in the space of the dynamical models of the chosen class. A number of distance functions for linear systems has been already introduced, in particular in the context of system identification: Martin's distance between cepstrums [1], subspace angles [2], gap metric [3] and its variants nu-gap [4] and graph metric [5], kernel methods [6]. Besides, a vast literature can be found about dissimilarity measures between hidden Markov models [7, 8], most of them variants [9] of the Kullback-Leibler divergence [10, 11]. However, a simple mental experiment is enough to understand how no single distance function can possibly outperform the others in each and every classification problem, since labels can be assigned arbitrarily to any given dataset.

The most reasonable thing to do, when possessing some a-priori information in terms of partially labelled data or similarity classes, is then try and *learn* in a supervised fashion the "best" distance function for a precise classification problem. This topic has become quite popular in the last few years (see for instance [12, 13, 14, 15, 16, 17, 18]). Many unsupervised algorithms, in particular, take an input dataset and embed it in some other space, implicitly learning a metric (locally linear embedding [19] among the others), but fail to learn a full metric for the whole input space.

In this paper we propose a differential-geometric method that, given a dataset of *unlabeled* linear systems of a given class, allows to learn the Riemannian metric that minimizes the inverse volume element around the data, this way forcing the geodesic of the manifold to pass through the region where the data are actually distributed. We will show how this improves the classification performance by posing the problem in the region where the systems actually live. More precisely, if the models belong to a Riemannian manifold $M$, any diffeomorphism of $M$ onto itself induces a *pullback metric* on $M$. By designing a suitable family of diffeomorphisms depending on a parameter $p$ we then obtain a family of pullback metrics on $M$. We can then analytically compute the volume element of the induced metric as a function of $p$, and pose a well-defined optimization problem to get the value of the parameter.

To prove the advantages of a classification scheme based on a metric learnt from the data, we consider two different problems: action recognition and identity recognition from gait. We use the image sequences collected in the Mobo database [20] to show experiments in which simple nearest-neighbor classifiers based on the pullback of the classical Fisher metric between stochastic models outperform analogous classifiers based on all the known distances between models.

## 2 Learning pullback metrics

The study of the geometrical structure of the space formed by a family of probability distribution is first due to Rao [21], and was developed by Nagaoka and Amari [22, 23]. A family $S$ of probability distributions $p(x,\xi)$ depending on a $n$-dimensional parameter $\xi$ can be regarded in fact as an $n$-dimensional manifold. If the Fisher information matrix

$$g_{ij} \doteq E\Big[\frac{\partial log p(x,\xi)}{\partial \xi_i}\frac{\partial log p(x,\xi)}{\partial \xi_j}\Big]$$

is nondegenerate, $G=[g_{ij}]$ is a Riemannian tensor, and $S$ is a Riemannian manifold. The squared distance $ds^2$ between two nearby distributions $p(x,\xi)$ and $p(x,\xi+\partial\xi)$ is known to be twice the respective *Kullback-Leibler divergence* [10]. Some approximations of this measure have been introduced based on Monte-Carlo methods or simplifying approximations [11, 7].
The problem of defining a metric for linear dynamical systems has been studied also in the framework of robust control. One criterion is to compare the outputs when the inputs are restricted to the class of inputs that give bounded outputs. This approach was introduced by Zames and El-Sakkary [3] [24] using the notion of *gap metric*. The original rationale for the (linear) gap metric was to provide a suitable topology in which small errors in the gap in open-loop systems would correspond to small errors in the norm of the stable closed-loop. Another solution called *$\nu$-gap metric* to the problem of comparing two systems that is appropriate for feedback analysis was suggested by Vinnicombe [4].
Another distance function has recently been introduced based on the notion of *cepstrum*, the inverse Fourier transform of the logarithm of the spectral density. Martin [1] defined a new metric for the set of SISO ARMA models, based on the inner product of the cepstra $(c_n)_{n\in\mathbb{Z}}$, $c_n^* = c_{-n}$, yielding a Euclidean metric $d(c,c') = |c-c'|^2 = \sum_n n|c_n - c'_n|^2$. De Cock [2] showed how the cepstrum norm of a SISO ARMA model is a function of the principal angles between the row space of the controllability matrices of the model and its inverse.

### 2.1 Learning metrics from the data: pullback metrics

None of the above distances (which have usually been designed for a precise goal) can obviosly outperform the others in every classification problem. Labels can be assigned arbitrarily to dynamical systems, so that a distance which works well for a particular labeling would fail if employed in a different classification task. When having some knowledge about the dataset (similarity between pairs, partial labeling, etc.) it makes much more sense to *learn* a distance metric reflecting this information, and use the learnt metric for classification (see [25]). Many unsupervised algorithms take an input dataset and embed it in some other space, implicitly learning a metric ([19] among the others). However, they fail to learn a full metric for the whole input space, so that the embedding has usually to be recomputed when new data are available[1].

Some simple notions of differential geometry can provide us with a fundamental tool to build a structured family of metrics on which to define an optimization problem, the basic ingredient of a metric learning algorithm. Consider a family of diffeomorphisms between the Riemannian manifold $M$ in which the dataset $D \subset M$ resides and itself:

$$F_\lambda : \quad M \quad \rightarrow \quad M, \quad m \quad \mapsto \quad F_\lambda(m) \tag{1}$$

This family of maps induces a family of *pullback* metrics on the space itself. Let us call $T_mM$ the tangent space to $M$ in $m$. Any diffeomorphism $F$ is associated with a *push-forward* map

$$\begin{array}{rccc} F_* : & T_mM & \rightarrow & T_{F(m)}M \\ & v \in T_mM & \mapsto & v' \in T_{F(m)}M \end{array}$$

defined as $(F_*v)f = v(f\circ F)$ for all the smooth functions $f$ on $M$. Then, given a Riemannian metric $g : TM \times TM \rightarrow \mathbb{R}$ on $M$, the diffeomorphism $F$ induces a *pullback* metric on $M$ as follows:

$$g_{*m}(u,v) \doteq g_{F(m)}(F_*u, F_*v). \tag{2}$$

Any parametric family of differentiable maps (1) generates naturally a parametric family of metrics on the original space $M$. The geodesic (shortest path) connecting two points under the pullback metric is the *lifting* of the geodesic associated with the original metric. In other words, the pullback geodesic between two points is just the geodesic connecting their images with respect to the original metric.

[1] Even though some work on out-of-sample extensions of spectral methods has been done [26].

# 3 Volume element minimization

As our task is to classify data, a natural optimization criterion would be to maximize the classification performance achieved by using the learnt metric for a simple nearest-neighbor classifier. Xing et al. [25] have recently proposed a way to solve this optimization problem for *linear* maps $y = A^{1/2}x$, when some pairs of points are known to be "similar". This leads to a parameterized family of Mahalanobis distances $\|x - y\|_A$. Given the linearity of the map the problem of minimizing the squared distance between similar points turns out to be convex, hence solvable with efficient numerical algorithms. Other people [27] have successfully faced the same problem in a linear framework, by solving an optimization problem based on information theory (*relevant component analysis*).

Unfortunately, maximizing classification performance is in practice very hard in a nonlinear context. A reasonable approach consists then on choosing a different objective function, which has to be correlated with the natural one in the sense that it should improve classification performances. An interesting choice has been recently suggested by G. Lebanon [28] in the context of document retrieval (see also [29]). The idea is to minimize the volume element associated with the pullback metric, in order to force the geodesics to pass through the most populated regions of the original space. More precisely, the function to minimize is

$$\mathcal{O}(D) = \prod_{k=1}^{N} \frac{(\det g(m_k))^{-\frac{1}{2}}}{\int_M (\det g(m))^{-\frac{1}{2}} dm}. \tag{3}$$

where $g(m_k)$ denotes the Riemannian metric at the point $m_k$ of a dataset living on a Riemannian manifold $M$. This is clearly related to finding a lower dimensional representation of the dataset, in a similar fashion to dimensionality reduction techniques like locally linear embedding [19] or laplacian eigenmaps [30][2]. The function (3) to optimize can also be seen as the inverse of Jeffreys' prior [31, 28].

## 3.1 Algorithm

To find the expression of the Gramian $\det g_*$ associated with the pullback metric (2) we first need to choose a base of the tangent space $T_mM$ to $M$. Let us then denote with $\{\partial_i\}$, $i = 1, ..., \dim M$ the base of $T_mM$. We can then compute the push-forward of the vectors of this base, getting in turn a base for $T_{F(m)}M$. By definition [32], the push-forward of a vector $v \in T_mM$ is

$$F_p(v) \doteq \tfrac{d}{dt} F_p(m + t \cdot v)\big|_{t=0}, \quad v \in T_mM. \tag{4}$$

The diffeomorphism $F_p$ then induces a base for the space of vector fields on $M$, $w_i \doteq \{F_p(\partial_i)\}$, for $i = 1, ..., \dim M$. We can rearrange these vectors as rows of a matrix $J = [w_1; \cdots; w_{\dim M}]$. The volume element of the pullback metric $g_*$ in a point $m \in M$ is then given by the determinant of the Gramian [28],

$$\det g_*(m) \doteq \det[g(F_{*p}(\partial_i), F_{*p}(\partial_j))]_{ij} = \det(J^T g J).$$

If $J$ is a square matrix (as in the following) we get simply

$$\det g_*(m) = \det(J)^2 \cdot \det g(m).$$

We can then easily find the expression of the objective function (3). Of course, once found the optimal metric, standard classification algorithms require to measure the geodesic distance between any two points. As geodesics for $g_{*p}$ are liftings, we only need to know the geodesics of the original Riemannian space.

# 4 Learning metrics for linear dynamical models

Even though large classes of non-linear models can be endowed with a Fisher metric tensor, most efforts have focused on the relatively simpler task of analyzing the Fisher geometry of some important classes of linear models [33, 34]. In particular, the analytic expressions of the entries of the Fisher information matrix for several manifolds of linear MIMO systems have been discovered by Hanzon et al. [35].

Let us consider first the class of stable autoregressive discrete-time processes of order 2, AR(2), in a stochastic setting in which as before the input signal is a Gaussian white noise with zero mean and unit variance. This set can be given a

[2] In [30] in particular, dimensionality reduction is considered a factor in improving classification.

Riemannian manifold structure under Fisher metric [36]. A natural parametrization uses the non-unit coefficients $(a_1, a_2)$ of the denominator of the transfer function, $h(z) = z^2/(z^2 + a_1 z + a_2)$ (which corresponds to the AR difference $y(k) = -a_1 y(k-1) - a_2 y(k-2)$). To impose stability the necessary conditions are

$$1 + a_1 + a_2 > 0, \quad 1 - a_1 + a_2 > 0, \quad 1 - a_2 > 0.$$

The manifold is then composed by a single connected component (see Figure 1). The Riemannian Fisher tensor can be expressed as [37]

$$g(a_1, a_2) = \frac{1}{(1 + a_1 + a_2)(1 - a_1 + a_2)(1 - a_2)} \begin{pmatrix} 1 + a_2 & a_1 \\ a_1 & 1 + a_2 \end{pmatrix}. \tag{5}$$

However, alternative local coordinates are given by the Schur parameters $\gamma_1 = a_1/(1 + a_2)$, $\gamma_2 = a_2$ under which the metric tensor simplifies as

$$g(\gamma_1, \gamma_2) = \frac{1}{(1 - \gamma_2^2)} \begin{pmatrix} \frac{(1+\gamma_2)^2}{(1-\gamma_1^2)} & 0 \\ 0 & 1 \end{pmatrix} \tag{6}$$

while the parameter domain becomes $|\gamma_1| < 1$, $|\gamma_2| < 1$.

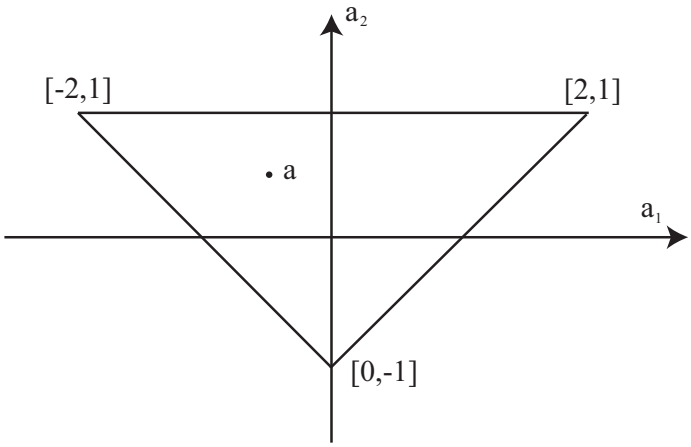


Figure 1: Graphical representation of the manifold of stable autoregressive systems of order 2, $AR(2)$, with the non-unit coefficients of the denominator of $h(z)$ as parameters. It forms a simplex with vertices $[-2, 1], [2, 1], [0, -1]$.

Consider now the class of stable discrete-time linear SISO systems of order 1,

$$\begin{cases} x(k+1) = ax(k) + bu(k) \\ y(k) = cx(k) \end{cases} \tag{7}$$

for which we can choose a canonical representation by setting $c = 1$, so that the transfer function is $h(z) = b/(z - a)$. We denote this class of systems with $M(1, 1, 1)$. Under the conditions $|a| < 1$ (stability) and $b \neq 0$ (minimality) this family of linear models form a manifold with local coordinates $(a, b)$, consisting of the two connected components $\{(a, b) : a < 0, b < 0\}$ and $\{(a, b) : a < 0, b > 0\}$. Using the generalized polar coordinates $r \doteq \frac{b}{\sqrt{1-a^2}}$, $\theta = \text{atanh}(a)$ the Fisher tensor becomes

$$g(r, \theta) = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \tag{8}$$

and the manifold $M(1, 1, 1)$ coincides with the two-dimensional plane, vertical axis excluded ($r \neq 0$).

## 4.1 Geodesics

We can use the analytic expressions of the Fisher metric to write the geodesic equations for these classes of linear systems. Using Einstein's notation the geodesic equation reads

$$\frac{d^2}{dt^2} x^a + \Gamma^a_{bc} \frac{d}{dt} x^b \frac{d}{dt} x^c = 0$$

where $\{\Gamma^a_{bc}\}$ are the Christoffel coefficients of the second kind [32], $\Gamma^a_{bc} = \Gamma_{bcd} g^{ad}$ where $g^{ad} = g^{-1}(a,d)$ are the entries of the inverse of the metric tensor, and the Christoffel coefficients of the first kind are a function of the derivatives of the metric:

$$\Gamma_{abc} \doteq \frac{1}{2} g^{ad} (\frac{d}{dx^c} g_{bd} + \frac{d}{dx^b} g_{cd} - \frac{d}{dx^d} g_{cb}).$$

Ravishanker [38] has explicitly computed the expression for the Fisher metric of ARMA(p,q) systems using as parametrization the inverses of poles and zeros, and we could then solve the geodesic differential equation analytically.

However, all the geodesics of stable AR(2) systems endowed with the Fisher metric (6) as a function of the Schur parameters have been analytically computed by Rijkeboer [39, 36].
The analytical expressions of the geodesics of the manifold $M(1,1,1)$ endowed with the Riemannian tensor induced by the scalar product of a Hilbert space $H_2$ have been instead computed by Hanzon [40] through an isometry with a Riemannian surface. We will use it instead of the Fisher tensor in the following[3]. More precisely, given two systems $\Sigma_1 = (r_1, \theta_2)$ and $\Sigma_2 = (r_2, \theta_2)$ their geodesic distance according to the metric (8) is

$$d(\Sigma_1, \Sigma_2) = (r_1^2 - 2 r_1 r_2 \cos(\theta_2 - \theta_1) + r_1^2)^{1/2}$$

if $|\theta_1 - \theta_2| \leq \pi$ and $sgn(r_1) = sgn(r_2)$;

$$d(\Sigma_1, \Sigma_2) = |r_1| + |r_2|$$

otherwise.

## 4.2 Volume elements of pullback metrics for spaces of linear models

The algorithm of Section 3.1 adapts easily to the case in which the dataset is a collection of linear systems.

One natural choice for a diffeomorphism of $AR(2)$ onto itself is suggested by the simplicial form of the manifold (see Figure 1),

$$F_p(\mathbf{m}) = F_p([m_1, m_2, m_3]) = \frac{1}{\lambda \cdot \mathbf{m}} [\lambda_1 \cdot m_1, \lambda_2 \cdot m_2, \lambda_3 \cdot m_3] \tag{9}$$

where $p = \lambda = [\lambda_1, \lambda_2, \lambda_3]$ with $\lambda_1 + \lambda_2 + \lambda_3 = 1$, while $\mathbf{m} = [m_1, m_2, m_3]$ collects the simplicial components of a system $\mathbf{a}$ in the manifold:

$$\mathbf{a} = [a_1, a_2]' = m_1 [0, -1]' + m_2 [2, 1]' + m_3 [-2, 1]', \quad \mathbf{a} \in AR(2). \tag{10}$$

**Theorem 1** *The volume element of the pullback metric on $AR(2)$ induced by the diffeomorphism (9) is given by*

$$det g_{*p}(m) \propto \frac{(\lambda_1 \lambda_2 \lambda_3)^2}{(\lambda \cdot \mathbf{m})^6} \cdot \frac{1}{m_1^2 m_2 m_3}.$$

Let us choose as base of the tangent space in $AR(2)$ the set $\partial_1 = [1/2, 1/2]'$, $\partial_2 = [-1/2, 1/2]'$. To compute the Gramian we need to express the diffeomorphism with respect to $\mathbf{a}$. From Equation (10) $a_1 = 2(m_2 - m_3)$, $a_2 = m_2 + m_3 - m_1$, while the inverse relation is $m_2 = \frac{1 + a_1 + a_2}{4}$, $m_3 = \frac{1 - a_1 + a_2}{4}$, $m_1 = \frac{1 - a_2}{2}$. Hence $F_\lambda$ can be expressed in terms of $a_1, a_2$ as

$$F_\lambda(\mathbf{a}) = \frac{1}{\Delta} [2\lambda_2(1 + a_1 + a_2) - 2\lambda_3(1 - a_1 + a_2), \lambda_2(1 + a_1 + a_2) + \lambda_3(1 - a_1 + a_2) - 2\lambda_1(1 - a_2)]'$$

where $\Delta = 2\lambda_1(1 - a_2) + \lambda_2(1 + a_1 + a_2) + \lambda_3(1 - a_1 + a_2)$, so that

$$\begin{aligned} F_\lambda(\mathbf{a} + t\mathbf{v}) = \quad & [2\lambda_2(1 + a_1 + tv_1 + a_2 + tv_2) - 2\lambda_3(1 - a_1 - tv_1 + a_2 + tv_2), \\ & \lambda_2(1 + a_1 + tv_1 + a_2 + tv_2) + \lambda_3(1 - a_1 - tv_1 + a_2 + tv_2) - 2\lambda_1(1 - a_2 - tv_2)]'. \end{aligned}$$

We can then compute[4] $\frac{d}{dt} F_\lambda(\mathbf{m} + t \cdot \mathbf{v})\big|_{t=0}$, and in particular

$$\begin{aligned} \mathbf{w}_1 = \quad \frac{d}{dt} F_\lambda(\mathbf{m} + t \cdot \partial_\mathbf{1})\big|_{t=0} = \quad & [2\lambda_1\lambda_2(3 - a_2 + a_1) + 2\lambda_3(2\lambda_2 - \lambda_1)(1 - a_1 + a_2), \\ & 2\lambda_1\lambda_2(3 - a_2 + a_1) + 2\lambda_1\lambda_3(1 - a_1 + a_2)] \end{aligned}$$

[3] The related geometries have indeed much in common, [40].

[4] The straightforward details are neglected to improve the readability of the proof.

while

$$\mathbf{w}_2 = \left.\frac{d}{dt}F_\lambda(\mathbf{m}+t\cdot\partial_{\mathbf{2}})\right|_{t=0} = [-2\lambda_1\lambda_3(3-a_2+a_1)+2\lambda_2(\lambda_1-2\lambda_3)(1+a_1+a_2),$$
$$2\lambda_1\lambda_3(3-a_2+a_1)+2\lambda_1\lambda_2(1+a_1+a_2)].$$

The determinant of the matrix $J$ is then (after a few passages)

$$\det J = 32\frac{\lambda_1\lambda_2\lambda_3}{\Delta^3} = \frac{1}{2}\frac{\lambda_1\lambda_2\lambda_3}{(\lambda\cdot\mathbf{m})^3}.$$

Eventually, the function (3) to maximize assumes the form

$$\mathcal{O}(p) = \prod_{k=1}^{N}\frac{(\lambda\cdot\mathbf{m}_k)^3}{\int_{AR(2)}(\lambda\cdot\mathbf{m})^3 m_1\sqrt{m_2m_3}\mathbf{dm}} \tag{11}$$

where the normalization factor

$$I(\lambda) = \int_{AR(2)}(\lambda\cdot\mathbf{m})^3 m_1\sqrt{m_2m_3}\mathbf{dm}$$

forbids trivial solutions in which the volume element is minimized by shrinking the whole space. It can be computed by decomposing the cube $(\lambda\cdot\mathbf{m})^3$ using Tartaglia's formula, obtaining

$$I(\lambda) = \sum_{c_1+c_2+c_3=3}\frac{3!}{c_1!c_2!c_3!}\prod_{j=1}^{3}\lambda_j^{c_j}\int_{AR(2)} m_1^{1+c_1}m_2^{1/2+c_2}m_3^{1/2+a_3}\mathbf{dm}.$$

About $M(1,1,1)$, Henon's map is the simplest non-trivial diffeomorphism of the plane: $x' = y+x^2+a$, $y' = -bx$. Here we choose a modified version:

$$F_p(r,\theta) = [\ b\theta + ar^2 + br,\quad br + a\theta^2 + b\theta\ ], \tag{12}$$

with $p = [a,b]$.

**Theorem 2** *The volume element of the pullback metric on $M(1,1,1)$ induced by the diffeomorphism (12) is given by*

$$detg_{*p}(x) = 4a^2(2ar\theta + br + b\theta)^2r^2.$$

It is easy to realize that in this case

$$J = \begin{bmatrix} 2ar+b & b \\ b & 2a\theta+b \end{bmatrix}$$

so that $\det J = 2a(2ar\theta + br + b\theta)$, while from Equation (8) $\det g(r,\theta) = r^2$.

This time the function (3) can be expressed as

$$\mathcal{O}(D) \propto \prod_k \frac{1}{r_k(2ar_k\theta_k) + br_k + b\theta_k}\cdot\frac{1}{I(a,b)} \tag{13}$$

where $I(a,b)$ is again the normalization factor. Note that when $I(p)$ is hard to compute, or the original manifold *has no finite volume* (and hence does not admit Jeffrey's prior), an acceptable replacement is $I(p) = \sum_{k=1}^{N}(detg(x))^{-1/2}$.

The functions (11) and (7) can be maximized by means of any numerical optimization scheme, simple (like gradient descent) or more sophisticated.

# 5 Human motion experiments

To test the effectiveness of the approach we considered a significant vision application, namely the recognition of actions and identities from image sequences. We used the Mobo database [20], a collection of 600 image sequences of 25 people walking on a treadmill in four different variants (slow walk, fast walk, walk on a slope, walk carrying a ball), seen from 6

different viewpoints corresponding to cameras equally distributed around the subject (see Figure 2-left). Each sequence of the database possesses three labels: action, view, and identity.

The database comes already with preprocessed silhouettes of the moving body as black and white images. As the results of the above sections concern only single input-single output linear systems, we chose to extract from each image the width of the minimum box containing the silhouette. This way, each image sequence is represented as a scalar signal. Each scalar sequence has been then passed as input to an identification algorithm: we hence generated a dataset of linear systems of the two classes discussed above, one system for each labeled motion sequence.

This metric learning approach is based on the assumption that the two optimization problems, i.e. maximizing classification performance and minimizing volume elements are indeed correlated. We empirically evaluated the goodness of this conjecture by measuring the performance of a nearest-neighbor classifier based on the optimal pullback metrics induced by the diffeomorphisms (12), (9), and comparing the results with those of NN classifiers based on all the other known distances (included the Fisher geodesic distance).

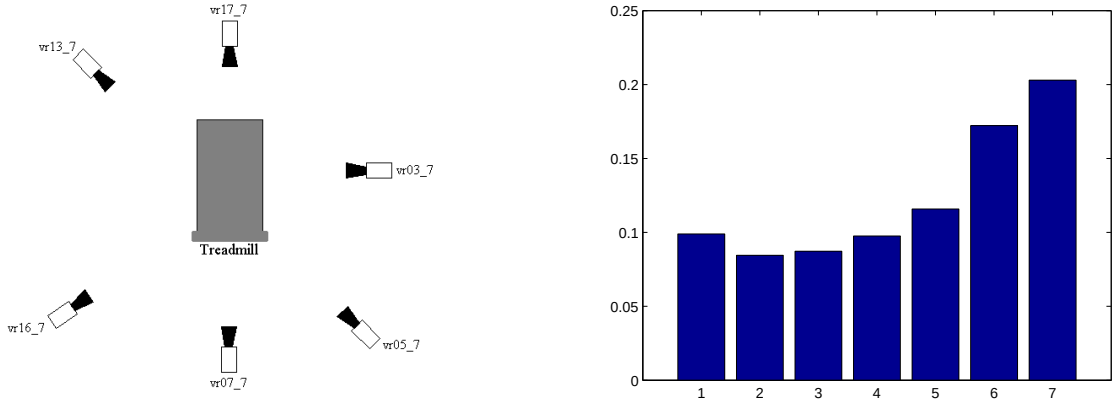


Figure 2: Left: location and orientation of the six cameras in the Mobo experiment (the origin of the frame is roughly in the position of the walking person on the treadmill). Right: Performance of NN classifier based on all the known distances between dynamical models, included the pullback metric induced by the Fisher geometry. Correct classification rate for the used metrics averaged on the size of the training set, in the ID experiment. 1 - Frobenius norm of the matrices in canonical form; 2 - gap metric; 3- $\nu$-gap matric; 4 - subspace angles; 5 - basis Fisher metric; 6 - pullback metric with diffeomorphism (14); 7 - pullback metric with diffeomorphism (9).

### 5.0.1 Identity recognition

In a first experiment we selected a training set of models, and used the pullback geodesic distance to classify the *identity* of the person appearing in a different set of randomly selected sequences. This is a very difficult problem, as there are 25 different people, and the one-dimensional signal we chose to represent sequences clearly provides insufficient information. However, measuring the comparative performance of the metrics can be useful to see how learning a metric appositively gives an advantages when using a NN classifier. We implemented a naive Frobenius norm of the system matrices in canonical form, gap and $\nu$-gap metrics [3, 4], subspace angles [2, 1], and of course the Fisher geodesic distance together with the associated optimal pullback metric.

Figure 2-right shows the percentage of correctly classified testing systems over several runs in which we randomly selected an increasing number of systems in both training and testing set. We also tested a different diffeomorphism for AR(2) systems, namely

$$F_\lambda(\mathbf{m}) = [\ \lambda m_1 + (1-\lambda)m_2, \quad \lambda m_2 + (1-\lambda)m_3, \quad \lambda m_3 + (1-\lambda)m_1\ ], \tag{14}$$

with $0 < \lambda < 1$. We can notice how the other distances do not show any significantly different behavior. This of course was to expect, since they have not been designed to solve classification tasks (usually for identification purposes). This has been the case in all our experiments. Hence, in the following we will only report the performance of the second best distance for comparison.

### 5.0.2 Action recognition

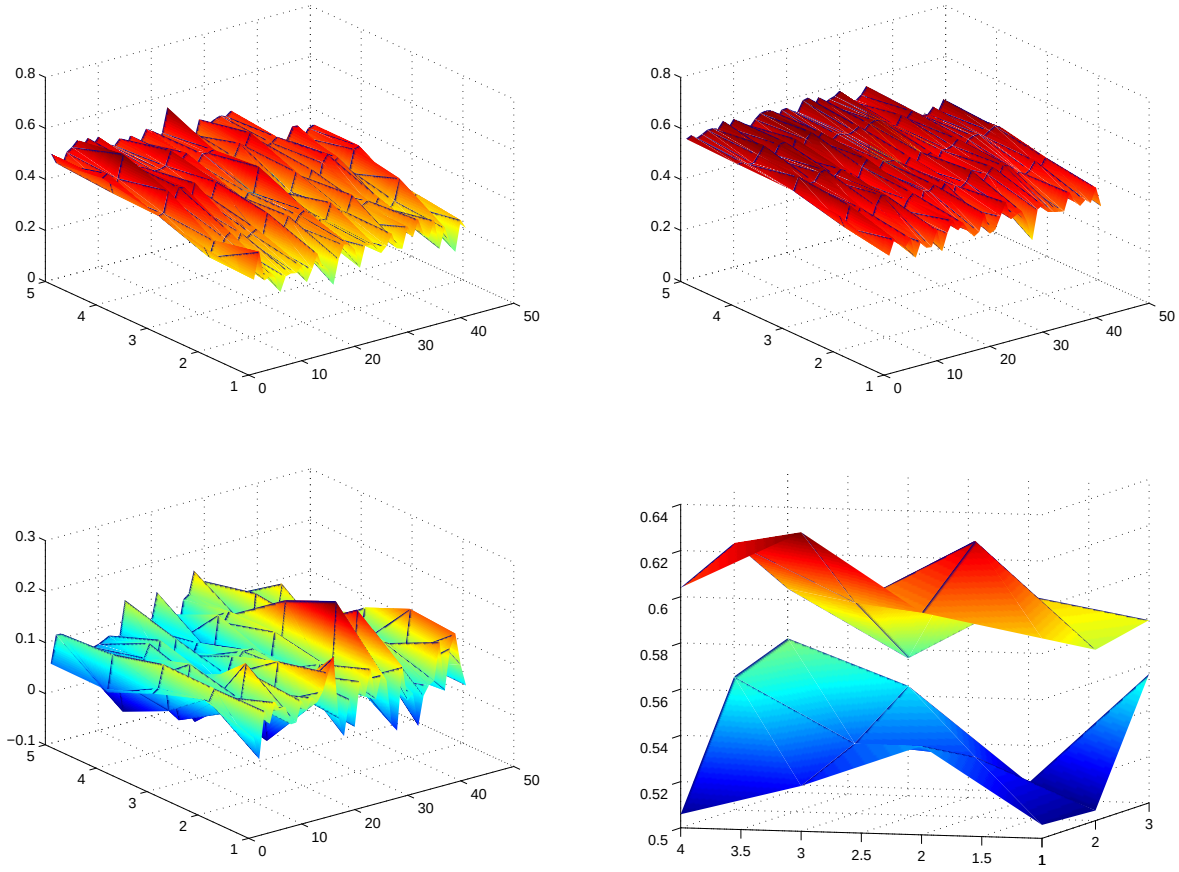


Figure 3: Top: Recognition performance of second-best distance (left) and optimal pullback metric for $AR(2)$ systems and diffeomorphism (14) (right) in the action recognition experiment. Here the whole dataset was considered, regardless of the viewpoint. Increasing $x$ indices $(1 : 50)$ mean decreasing size of the testing set. Increasing $y$ indices $(1 : 5)$ represent increasing size of the training set. Bottom-left: Difference between the two performances. Bottom-right: relative performances for view 5 only. Again, the pullback metric outperforms the best competitor.

In a second experiment, we selected a training set of models, and used the pullback geodesic distance to classify the *action* performed by the person in a different set of randomly selected sequences. Recall that all the actions in the Mobo database are just slightly different variations of the walking gait, and we were using a *single* scalar feature. Figure 3 illustrates the relative performances of the optimal pullback metric for autoregressive systems under deformation map (14) and its best competitor. The correct classification rate has been measured for different sizes of both training and testing set, by randomly selecting for each class of action an equal increasing number of systems to build the two collections. To test the approach more thoroughly we also conducted six separate experiments by selecting the portion of the dataset associated with a single view, for each possible view. These are completely different classification problems, as the datasets can turn out to live in different submanifolds. It is in fact interesting to try and see what happens when we choose a different class of diffeomorphisms. Figure 4 illustrates the average performance of the classifiers associated with the optimal pullback metric for AR(2) models with (this time) diffeomorphism (9) and the best competing distance for four view-dependent experiments. This time increasing abscissae (from 1 to 7) mean an increased testing set size. As usual the average is computed by repeated random selections, and the optimal metric performs far better than the others.

Of course the choice of the family of deformations affects the range of candidate metric tensors the algorithm can select from. We have seen as a matter of fact how the first family of mappings, which depends on 2 parameters instead of 1,

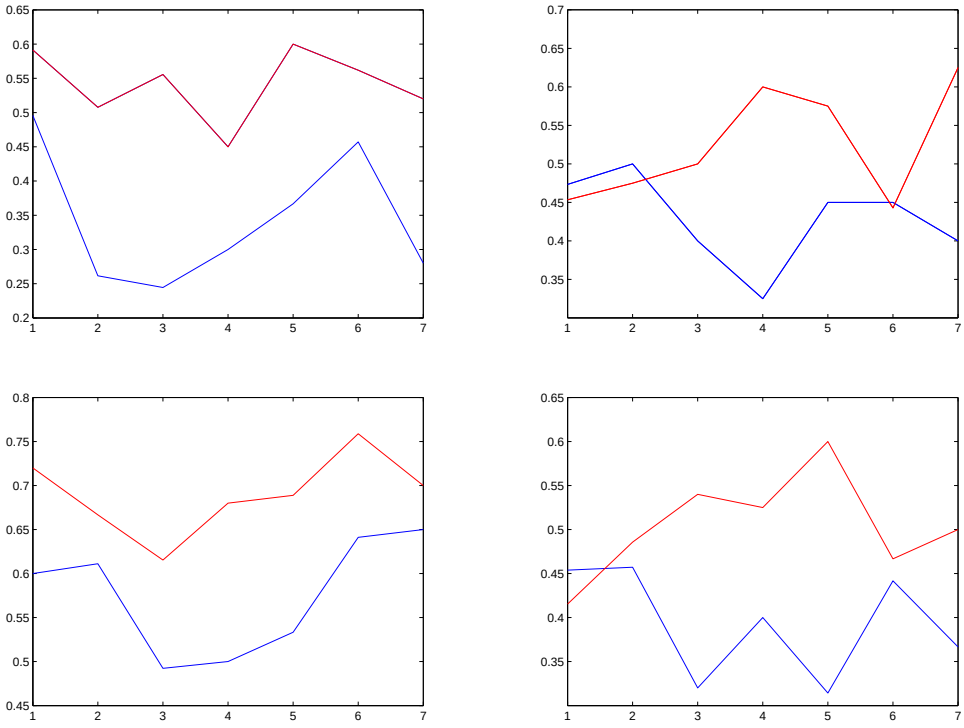


Figure 4: Correct action classification rates for four different viewpoints. Sequences are represented as autoregressive order 2 systems. The performance (in red) of the pullback metric associated with the map (9) is shown outperforming the best competitor (in blue) for increasing sizes of the testing set.

produces a much wider selection choice for the optimization algorithm, eventually delivering a better performance. The same phenomenon can be appreciated in the identity experiment (see Figure 2).

It is really interesting to understand whether the size of the training set we use to learn the optimal metric has or not an influence on the classification performance. Of course, as our method exploits the training set to understand the local structure of the data, it is reasonable to conjecture that larger training sets should lead to better recognition rates, as long as the training data better represent the whole dataset. This is confirmed by Figure 5, where the recognition rate is plotted against the size of the training set. Finally, let us visualize the effect of the optimal diffeomorphism on the dataset. We can easily do it in the action recognition experiment, as there are only 4 labels. Figure 6-left shows an instance of the original dataset represented as a set of points in the $AR(2)$ simplex. Points of the training set are drawn as squares, while systems in the testing set are represented as crosses. Figure 6-right shows how the diffeomorphism deforms the space, plotting the images of the dataset according to the optimal map. Data-points are colored according to their true class. We can visually appreciate how the deformation improves the separation between clusters.

# 6 Conclusions

In this paper we discussed the problem of learning the "best" metric for a classification problem involving linear dynamical models. Given a training set of models we pose a related optimization problem in which the pullback metric induced by a diffeomorphism which minimizes the volume element around the available data is learned. We adopt as basis metric tensor the classical Fisher information matrix. This yields a global embedding, while usual spectral methods only provide images of training points.

We have also shown experimental results concerning identity and action recognition, which proved how such a learnt metric actually improves classification performance with respect to the Fisher metric and all the other known distances between dynamical models. Setting and solving the minimization problem for MIMO systems is the next natural step. This

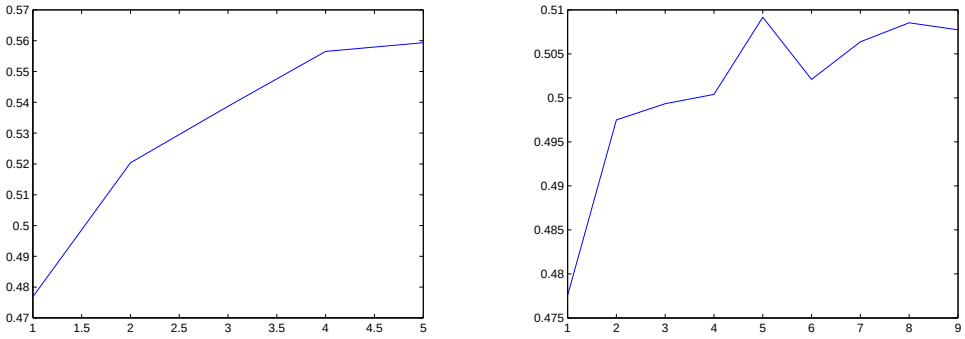


Figure 5: Effect of the size of the training set on the recognition accuracy in the action recognition experiment, $M(1,1,1)$ systems. As the size of $D$ increases (abscissa indices) the number of correct matches grows in both the overall experiment for all view (left) and for view 2 (right).

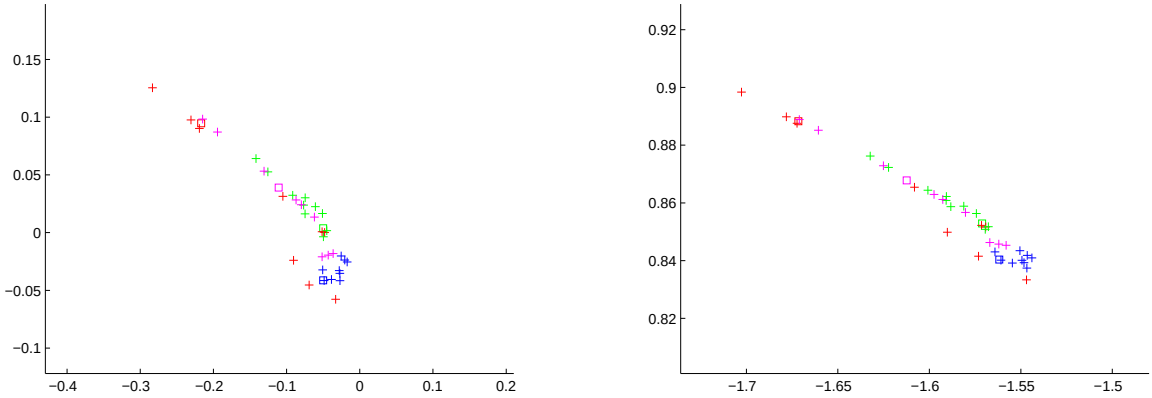


Figure 6: Effect of the optimal diffeomorphism (9) on the dataset.

will be of course of paramount importance to test this approach in a more realistic scenario.

# References

[1] Martin, R.J.: A metric for ARMA processes. IEEE Transactions on Signal Processing **48(4)** (April 2000) 1164–1170 1, 2, 7

[2] Cock, K.D., Moor, B.D.: Subspace angles and distances between arma models. Systems and Control Letters (2002) 1, 2, 7

[3] Zames, G., El-Sakkary, A.K.: Unstable systems and feedback: The gap metric. In: Proc. 18th Allerton Conference on Communications, Control, and Computers. (Urbana, IL, October 1980) 380–385 1, 2, 7

[4] Vinnicombe, G.: A $\nu$-gap distance for uncertain and nonlinear systems. In: Proceedings of the 38th IEEE CDC. (Phoenix, AZ, 1999) 1, 2, 7

[5] Vidyasagar, M.: The graph metric for unstable plants and robustness estimates for feedback stability. AC **29** (1984) 403–417 1

[6] Smola, A., Vishwanathan, S.: Hilbert space embeddings in dynamical systems. In: Proceedings of the 13th IFAC symposium on system identification. (August 2003) 760 – 767 1

[7] Silva, J., Narayanan, S.: Average divergence distance as a statistical discrimination measure for hidden Markov models. Submitted to IEEE Transactions on Speech and Audio Processing (2004) 1, 2

[8] Lyngso, R.B., Pedersen, C.N.S., Nielsen, H.: Metrics and similarity measures for hidden Markov models. In: Proceedings of ISMB 1999, AAAI Press. (1999) 178–186 1

[9] Do, M.N.: Fast approximation of Kullback - Leibler distance for dependence trees and hidden Markov models. IEEE Signal Processing Letters **10(4)** (April 2003) 115–118 1

[10] Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics **22** (1951) 79–86 1, 2

[11] Juang, B.H., Rabiner, L.: A probabilistic distance measure for hidden Markov models. ATT Technical Journal **64(2)** (1985) 391–408 1, 2

[12] Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning distance functions using equivalence relations. In: ICML03. (2003) 11–18 1

[13] Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning distance functions for image retrieval. In: CVPR04. Volume 2. (2004) 570–577 1

[14] Bilenko, M., Basu, S., Mooney, R.: Integrating constraints and metric learning in semi-supervised clustering. In: Proc. of 21st International Conference on Machine Learning (ICML'04). (2004) 1

[15] Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. In: Advances in Neural Information Processing Systems. (2004) 1

[16] Tsang, I., Kwok, J., Bay, C., Kong, H.: Distance metric learning with kernels. In: Proceedings of the International Conference on Artificial Intelligence. (2003) 1

[17] Zhang, Z.: Learning metrics via discriminant kernels and multidimensional scaling: Toward expected euclidean representation. In: ICML'03. (Hong Kong, 2003) 1

[18] Eick, C., Rouhana, A., Bagherjeiran, A., Vilalta, R.: Using clustering to learn distance functions for supervised similarity assessment. In: International Conference on Machine Learning and Data Mining. (2005) 1

[19] Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290(5500)** (2000) 23232326 1, 2, 3

[20] Gross, R., Shi, J.: The cmu motion of body (mobo) database. Technical report, Carnegie Mellon University (2001) 1, 6

[21] Rao, C.: Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37** (1945) 81–91 2

[22] Amari, S.I.: Differential geometry of curved exponential families: curvatures and information loss. Ann. Statist. **10** (1982) 357–87 2

[23] Amari, S.I.: Differential geometric methods in statistics. Springer-Verlag, Berlin (1985) 2

[24] El-Sakkary, A.K.: The gap metric: Robustness of stabilization of feedback systems. AC **26** (1985) 240–247 2

[25] Xing, E., Ng, A., Jordan, M., Russel, S.: Distance metric learning with applications to clustering with side information. In: Advances in Neural Information Processing Systems, 15. The MIT Press (2003) 2, 3

[26] Bengio, Y., Paiement, J.F., Vincent, P.: Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering. Technical report, Universite' de Montreal (2003) 2

[27] Shental, N., Hertz, T., Weinshall, D., Pavel, M.: Adjustment learning and relevant component analysis. In: ECCV'02. (2002) 3

[28] Lebanon, G.: Computing the volume element of a family of metrics on the multinomial simplex. Technical report, CMU-CS-03-145, Carnegie Mellon University (2003) 3

[29] Murray, M., Rice, J.: Differential Geometry and Statistics. CRC Press (1993) 3

[30] Belkin, M., Niyogi, P.: Semi-supervised learning on riemannian manifolds. Machine Learning **56** (2004) 209239 3

[31] Jeffreys, H.: Theory of Probability. Clarendon Press, New York (1983) 3

[32] Petersen, P.: Riemannian Geometry. Springer-Verlag, (Berlin) 3, 5

[33] Hanzon, B., Peeters, R.: Aspects of Fisher geometry for stochastic linear systems, problem 25. In Blondel, V., Megretsky, A., eds.: Open Problems in Mathematical Systems and Control Theory. MTNS2002 (2002) 27–30 3

[34] Peeters, R.: System identification based on Riemannian geometry. PhD dissertation, Tinbergen Inst. Res. Ser., Amsterdam (1994) 3

[35] Peeters, R., Hanzon, B.: Symbolic computation of Fisher information matrices for parameterized state-space systems. Automatica **35** (1999) 1059–1071 3

[36] Rijkeboer, A.: Differential geometric models for time-varying coefficients of autoregressive processes. PhD thesis, Tilburg University, Tilburg (1994) 4, 5

[37] Peeters, R., Hanzon, B.: On the Riemannian manifold structure of classes of linear systems. In: Equadiff2003. (2003) 4

[38] Ravishanker, N., Melnick, E., Tsai, C.L.: Differential geometry of ARMA models. J. Time Series Anal. **11** (1990) 259–274 5

[39] Rijkeboer, A.: Fisher optimal approximation of an AR(n)-process by an AR(n-1)-process. In Nieuwenhuis, J., Praagman, C., Trentelman, H., eds.: Proceedings of ECC'93. (Groningen, 1993) 1225–1229 5

[40] Hanzon, B.: Identifiability, recursive identification and spaces of linear dynamical systems. CWI Tracts **63-64** (Amsterdam 1989) 5