

# Learning Pullback Metrics for Linear Models

Fabio Cuzzolin  
fabiocuzzolin@gmail.com

Oxford Brookes University

**Abstract.** In this paper we present an unsupervised differential-geometric approach for learning Riemannian metrics for dynamical models. 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. The problem of classifying motions, encoded as dynamical models of a certain class, can then be posed on the learnt manifold. 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

Manifold learning [1–6] has become a popular topic in machine learning and computer vision in the last few years, as many objects of interests (like natural images, or sequences representing walking persons), in spite of their apparent high dimensionality, live in a non-linear space of reduced dimension. Many unsupervised algorithms (e.g. locally linear embedding [7]) take an input dataset and embed it into some other space, implicitly learning a metric. However, they fail to learn a full metric for the whole input space (even though extensions which take this issue into account have been recently formulated [8]). On the other side, approaches to successfully reduce metric learning to constrained least square optimization in the linear case have been proposed [9, 10].

In particular, it makes a lot of sense to represent videos or image sequences as realizations of some sort of dynamical model, either stochastic (e.g. HMMs) or deterministic (e.g. ARMA). This approach already proved to be effective in contexts like video coding (e.g. dynamic textures [11]). A number of metrics or distance functions for linear systems has been already introduced, in particular in the context of system identification: cepstrum distances [12], subspace angles [13], gap metric [14] and its variants nu-gap [15] and graph metric [16], kernel methods [17]. Besides, a vast literature exists about dissimilarity measures between Markov models, variants of the Kullback-Leibler divergence [18].

Consider though the problem of classifying a dynamical model (as representative of an input image sequence). A simple mental experiment is enough to understand that no single distance function can possibly outperform all the others in each and every classification problem, since data-points can be endowed with different labeling while maintaining the same geometrical structure. In gait

analysis, for instance, each image sequence possesses several different labels: the identity of the moving person, the action performed, the viewpoint (when several cameras are presented [19]), the emotional status, etc.

The most reasonable thing to do when possessing some a-priori information is then try and *learn* in a supervised fashion the “best” distance function for a specific classification problem. As even linear dynamical models live in a non-linear space, the need for a principled way of learning Riemannian metrics from the data naturally arises. An interesting tool is provided by the formalism of *pullback metrics*. If the models belong to a Riemannian manifold  $M$ , any diffeomorphism of  $M$  onto itself induces such a 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 optimize on.

In this paper we propose to apply this formalism to the case of linear dynamical models endowed with the classical Fisher metric. Analytical expressions of Fisher metric and related geodesics for several classes of linear dynamical systems have been provided by Hanzon [20], Ravishanker [21], and Rijkeboer [22]. Given a dataset of dynamical models a “natural” Riemannian metric is learned by inverse volume element minimization on a family of pullback metrics induced by a class of diffeomorphisms. The learnt metric can later be used in any classification scheme. We argue that this improves the classification performance by posing the problem in the region where the systems actually live.

We apply this learning scheme to action recognition and identity recognition from gait. We use the image sequences collected in the Moby database [19] to show experiments in which (as a reference) simple nearest-neighbor classifiers based on pullback Fisher metric between stochastic models outperform analogous classifiers based on classical a-priori distances.

## 2 Learning non-linear manifolds of dynamical models

**Metric and manifold learning.** Learning distance functions or metrics to improve, for instance, classification performance is a sensible approach when some a-priori information is available [1–6]. A natural optimization criterion consists on maximizing the classification performance achieved by using the learnt metric in a simple nearest-neighbor classifier. Xing et al. [9] 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. Shental et al. [10] have successfully faced the same problem in a linear framework, by posing an optimization problem based on information theory (*relevant component analysis*). However, in several real-world applications the data are inherently nonlinear. This is exactly the case of dynamical models (and human motions, in particular). The need for a way of learning Riemannian metrics from such data arises.

**Pullback metrics.** Some notions of differential geometry provide us indeed with a tool for building a structured family of metrics on which to define an optimization problem, the basic ingredient of a metric learning algorithm. The idea is the following. Suppose your dataset already lives on a Riemannian manifold  $M$  of some sort: a Riemannian metric is defined in any point of the manifold. Any diffeomorphism (a differentiable map) from  $M$  to itself induces a new metric, called “pullback metric” (see below). If we manage to define an entire class of diffeomorphisms depending on some parameter  $\lambda$  we then get the corresponding family of pullback metrics, also depending on  $\lambda$ .

We can then define an optimization problem over this family of metrics in order to select an “optimal” metric, which in turn determines the desired manifold. Of course, the nature of this manifold depends on the objective function we choose to optimize. Following Lebanon [23] we propose to maximize the inverse volume of the space around the dataset to select the manifold on which the latter lives.

**Proposed methodology.** In this paper the dataset is composed by dynamical models representing motions. Many classes of dynamical systems do live in a Riemannian manifold  $M$  associated with the Fisher information metric [24]. We can then apply the formalism of pullback metrics, and learn a manifold associated with a training set  $D$  of movements. When new motions are acquired they can then be classified on this reduced space within a three-step procedure:

1. each sequence of measurements representing a dynamical data-set is mapped into the parameters of a dynamical model describing the sequence by parameter identification;
2. a parametric metric is learned for the resulting set of dynamical systems;
3. standard classifiers (e.g. k-nearest neighbor, SVM, etc.) are used to classify the systems according to the new metric.

### 3 Volume minimization for pullback metric learning

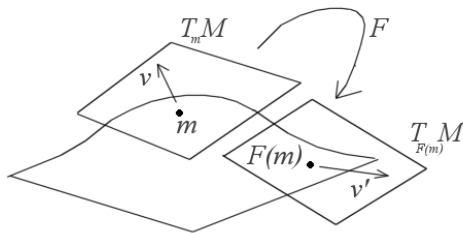
Consider a family of diffeomorphisms between the Riemannian manifold  $M$  in which the dataset  $D = \{m_1, \dots, m_N\} \subset M$  resides and itself:  $F_p : M \rightarrow M$ ,  $m \mapsto F_p(m)$ ,  $m \in M$ . Let us denote by  $T_m M$  the tangent space to  $M$  in  $m$ . Any such diffeomorphism  $F$  is associated with a *push-forward* map

$$\begin{aligned} F_* : T_m M &\rightarrow T_{F(m)} M \\ v \in T_m M &\mapsto F_* v \in T_{F(m)} M \end{aligned} \quad (1)$$

defined as  $F_* v(f) = v(f \circ F)$  for all the smooth functions  $f$  on  $M$  (see Figure 1). Consider now a Riemannian metric  $g : TM \times TM \rightarrow \mathbb{R}$  on  $M$ . Roughly speaking, a Riemannian metric determines how to compute scalar products of tangent vectors  $v \in T_m M$ . The map  $F$  induces a *pullback* metric on  $M$  as

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

Scalar products of two vectors  $u, v$  according to  $g_{*m}$  are computed as the scalar product with respect to the original metric  $g$  of the pair of vectors  $F_* u, F_* v$



**Fig. 1.** The push-forward map associated with a diffeomorphism on a Riemannian manifold  $M$ .

which are mapped onto  $u, v$  by  $F$ .

Any parametric family of differentiable maps naturally generates an entire parametric family of metrics on  $M$ . The pullback geodesic between two points is then the geodesic connecting their images with respect to the original metric.

**Volume element minimization.** From this parameterized family of metrics, we wish to determine the Riemannian manifold which “best” represents the dataset, according to some criterion. We propose here to *minimize the volume element* associated with a metric, as suggested by G. Lebanon [23] in the context of document retrieval (see also [25]). Equivalently, we seek to maximize

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

where  $g(m_k)$  denotes the Riemannian metric in the point  $m_k$  of the data-set  $D$  living on a Riemannian manifold  $M$ . This amounts to finding a lower dimensional representation of the dataset, in a similar fashion to locally linear embedding [7] or laplacian eigenmaps [26], where dimensionality reduction is often considered a factor in improving classification.

**Computing the Gramian.** The computation of (3) requires that of the *Gramian*  $\det g$ . To find the expression of the Gramian associated with the pullback metric (2) we first need to choose a base of the tangent space  $T_m M$  to  $M$ . Let us then denote with  $\{\partial_i\}$ ,  $i = 1, \dots, \dim M$  the base of  $T_m M$ .

We can compute the push-forward of the vectors of this base, getting in turn a base for  $T_{F(m)} M$ . By definition, the push-forward of a vector  $v \in T_m M$  is [23]

$$F_p(v) \doteq \frac{d}{dt} F_p(m + t \cdot v) \Big|_{t=0}, v \in T_m M. \quad (4)$$

The diffeomorphism  $F_p$  induces a base for the space of vector fields on  $M$ ,  $w_i \doteq \{F_p(\partial_i)\}$ , for  $i = 1, \dots, \dim M$ . We can rearrange these vectors as rows of a matrix

$$J = [w_1; \dots; w_{\dim M}]. \quad (5)$$

The volume element of the pullback metric  $g_*$  in a point  $m \in M$  is then the determinant of the Gramian [23]:  $\det g_*(m) \doteq \det[g(F_{*p}(\partial_i), F_{*p}(\partial_j))]_{ij} = \det(J^T g J)$ .

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

$$\det g_*(m) = \det(J)^2 \cdot \det g(m). \quad (6)$$

Plugging (6) in (3) we obtain the function to minimize.

**Algorithm.** We have then a method to select an optimal metric for a dataset  $D = \{m_1, \dots, m_N\}$  of points of a Riemannian manifold  $M$  with (basis) metric  $g$ .

1. First, a family  $\{F_p, p \in P\}$  of diffeomorphisms ( $P$  the parameter space) from  $M$  to itself is designed to provide a large search space of metrics (the variable in our optimization scheme) from which to select the optimal one;
2.  $F_p$  induces a family  $g_{*p}$  of pullback metrics (2) on  $M$ ;
3. we can then pose an optimization problem  $\max_{p \in P} \mathcal{O}(p)$ , where

$$\mathcal{O}(p) = \prod_{k=1}^N \frac{(\det g_{*p}(m_k))^{-\frac{1}{2}}}{\int_M (\det g_{*p}(m))^{-\frac{1}{2}} dm} \quad (7)$$

is the objective function, the inverse volume of the manifold in the neighborhood of the  $N$  points  $\{m_k, k = 1, \dots, N\}$  of the dataset  $D$ ;

4. this yields an optimal pullback metric  $\hat{g}_*$ ;
5. knowing the geodesics of  $M$  we can compute the distance between two points according to  $\hat{g}_*$ ;
6. this distance function can then be used to cluster or classify the dataset.

In the following Section we will show how to apply this optimization scheme to the case of linear dynamical models, endowed with the Fisher metric. In particular we will consider the simple case of autoregressive models of order 2, for which both Fisher metric and geodesics are analytically known, and recover closed-form expressions for the objective function (7).

## 4 Learning pullback metrics for linear models

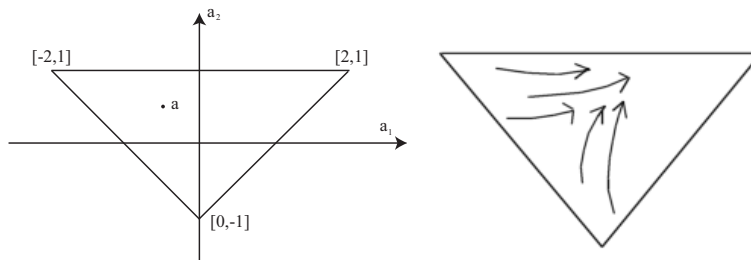
As linear dynamical models live in a Riemannian space, we can apply to them the pullback metric formalism and obtain a family of metrics on which to optimize.

**Fisher metric.** The study of the geometrical structure of the space formed by a family of probability distribution was first due to Rao, and was developed by Nagaoka and Amari [24]. 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

$$g_{ij} \doteq E \left[ \frac{\partial \log p(x, \xi)}{\partial \xi_i} \frac{\partial \log p(x, \xi)}{\partial \xi_j} \right] \quad (8)$$

is nondegenerate,  $G = [g_{ij}]$  is a Riemannian tensor, and  $S$  is a Riemannian manifold. In particular, the analytic expressions of the entries of the Fisher information matrix for several manifolds of linear MIMO systems have been obtained by Hanzon et al. [27].

**An interesting class of linear dynamical models.** Let us consider in particular the class of stable autoregressive discrete-time processes of order 2,  $AR(2)$ , in a stochastic setting in which the input signal is a Gaussian white noise with zero mean and unit variance. This set can be given a Riemannian manifold structure under Fisher metric. A natural parametrization uses the non-unit coefficients  $(a_1, a_2)$  of the denominator of the transfer function,  $h(z) = z^2/(z^2 + a_1z + a_2)$  (which corresponds to the AR difference  $y(k) = -a_1y(k-1) - a_2y(k-2)$ ). To impose stability the necessary conditions are  $1 + a_1 + a_2 > 0$ ,  $1 - a_1 + a_2 > 0$ , and  $1 - a_2 > 0$ . The manifold is then composed by a single



**Fig. 2.** Left: 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  $[-2, 1], [2, 1], [0, -1]$ . Right: Effect of a diffeomorphism of the form (10) on the  $AR(2)$  simplex.

connected component (see Figure 2-left). The Riemannian Fisher tensor can be expressed, in alternative local coordinates given by the Schur parameters  $\gamma_1 = a_1/(1 + a_2)$ ,  $\gamma_2 = a_2$ , as [28]

$$g(\gamma_1, \gamma_2) = \frac{1}{(1 - \gamma_2^2)} \begin{pmatrix} \frac{(1+\gamma_2)^2}{(1-\gamma_1^2)} & 0 \\ 0 & 1 \end{pmatrix}. \quad (9)$$

**Geodesics.** To compute the distance between two points of a Riemannian manifold (and in particular two dynamical models) it is not sufficient to know the metric: it is necessary to compute (analytically or numerically) the shortest path connecting them on the manifold (geodesic). All the geodesics of stable  $AR(2)$  systems endowed with the Fisher metric (9) as a function of the Schur parameters have been analytically computed by Rijkeboer [22].

**A family of diffeomorphisms for  $AR(2)$ .** To build a parameterized family of Riemannian metrics for  $AR(2)$  we can apply the optimal pullback metric scheme of Section 2 to the case in which the dataset  $D$  is a collection of linear systems which leaves in a Riemannian manifold  $M = AR(2)$ , the triangle of Figure 2. It is then necessary to choose a family of diffeomorphisms of  $M$  onto itself. Clearly the choice of a class of diffeomorphisms depends on the class of dynamical systems we adopt, but it is not univocal. The more sophisticated the

set of diffeomorphisms, the larger is the search space to optimize the metric on. Some hints can be provided by the geometry itself of the space  $M$  of dynamical models. As a matter of fact one possible choice for a diffeomorphism of  $AR(2)$  onto itself is suggested by the triangular form of the manifold (see Figure 2),

$$F_p(\mathbf{m}) = F_p([m_1, m_2, m_3]) = \frac{[\lambda_1 m_1, \lambda_2 m_2, \lambda_3 m_3]}{\lambda \cdot \mathbf{m}} \quad (10)$$

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 coordinates of a system  $\mathbf{a} \in AR(2)$  in the manifold:

$$\mathbf{a} = [a_1, a_2]' = m_1[0, -1]' + m_2[2, 1]' + m_3[-2, 1]' \quad (11)$$

and  $\lambda \cdot \mathbf{m}$  denotes the scalar product of the two vectors. The application (10) stretches the triangle towards the vertex with the largest  $\lambda_i$  (Figure 2-right).

**Volume element for  $AR(2)$ .** It is then possible to apply the method of Section 3 to find the analytical expression of the determinant of the Gramian  $detg_{*\lambda}(\lambda, \mathbf{m})$  (6), as a function of the parameter  $\lambda$  of the family of diffeomorphisms (10). By plugging it in the expression for the inverse volume (3) we finally obtain the objective function  $\mathcal{O}(\lambda)$  to optimize to find an optimal metric.

**Theorem 1.** *The volume element of the pullback metric on  $AR(2)$  induced by the diffeomorphism (10) is given by*

$$detg_{*\lambda}(\lambda, \mathbf{m}) \propto \frac{(\lambda_1 \lambda_2 \lambda_3)^2}{(\lambda \cdot \mathbf{m})^6} \cdot \frac{1}{m_1^2 m_2 m_3}. \quad (12)$$

*Proof.* 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 (11)  $a_1 = 2(m_2 - m_3)$ ,  $a_2 = m_2 + m_3 - m_1$ , 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

$$F_\lambda(\mathbf{a}) = \frac{1}{\Delta} \begin{bmatrix} 2\lambda_2(1 + a_1 + a_2) - 2\lambda_3(1 - a_1 + a_2), \\ \lambda_2(1 + a_1 + a_2) + \lambda_3(1 - a_1 + a_2) - 2\lambda_1(1 - a_2) \end{bmatrix}' \quad (13)$$

where  $\Delta = 2\lambda_1(1 - a_2) + \lambda_2(1 + a_1 + a_2) + \lambda_3(1 - a_1 + a_2)$ , so that  $F_\lambda(\mathbf{a} + t\mathbf{v}) = [2\lambda_2(1 + a_1 + tv_1 + a_2 + tv_2) - 2\lambda_3(1 - a_1 - tv_1 + a_2 + tv_2), \lambda_2(1 + a_1 + tv_1 + a_2 + tv_2) + \lambda_3(1 - a_1 - tv_1 + a_2 + tv_2) - 2\lambda_1(1 - a_2 - tv_2)]'$ . We can then compute<sup>4</sup>  $\frac{d}{dt}F_\lambda(\mathbf{m} + t \cdot \mathbf{v})|_{t=0}$ , and in particular

$$\mathbf{w}_1 = \frac{d}{dt}F_\lambda(\mathbf{m} + t\partial_1)|_{t=0} = \begin{bmatrix} 2\lambda_1\lambda_2(3 - a_2 + a_1) + 2\lambda_3(2\lambda_2 - \lambda_1)(1 - a_1 + a_2), \\ 2\lambda_1\lambda_2(3 - a_2 + a_1) + 2\lambda_1\lambda_3(1 - a_1 + a_2) \end{bmatrix} \quad (14)$$

while

$$\mathbf{w}_2 = \frac{d}{dt}F_\lambda(\mathbf{m} + t\partial_2)|_{t=0} = \begin{bmatrix} -2\lambda_1\lambda_3(3 - a_2 + a_1) + 2\lambda_2(\lambda_1 - 2\lambda_3)(1 + a_1 + a_2), \\ 2\lambda_1\lambda_3(3 - a_2 + a_1) + 2\lambda_1\lambda_2(1 + a_1 + a_2) \end{bmatrix}. \quad (15)$$

<sup>4</sup> The straightforward details are not reported to improve the readability of the proof.

The determinant of the matrix  $J = [\mathbf{w}_1; \mathbf{w}_2]$  (5) is then (after a few passages)

$$\det J = 32 \frac{\lambda_1 \lambda_2 \lambda_3}{\Delta^3} = \frac{1}{2} \frac{\lambda_1 \lambda_2 \lambda_3}{(\lambda \cdot \mathbf{m})^3}. \quad \square \quad (16)$$

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

$$\mathcal{O}(\lambda) = \prod_{k=1}^N \frac{(\lambda \cdot \mathbf{m}_k)^3}{\int_{AR(2)} (\lambda \cdot \mathbf{m})^3 m_1 \sqrt{m_2 m_3} \mathbf{d}\mathbf{m}} \quad (17)$$

where the normalization factor  $I(\lambda) = \int_{AR(2)} (\lambda \cdot \mathbf{m})^3 m_1 \sqrt{m_2 m_3} \mathbf{d}\mathbf{m}$  forbids trivial solutions in which the volume element is minimized by shrinking the whole space. It can be computed by decomposing  $(\lambda \cdot \mathbf{m})^3$  using Tartaglia's formula:

$$I(\lambda) = \sum_{c_1+c_2+c_3=3} \frac{3!}{c_1!c_2!c_3!} \cdot \prod_{j=1}^3 \lambda_j^{c_j} \int_{AR(2)} m_1^{1+c_1} m_2^{1/2+c_2} m_3^{1/2+c_3} \mathbf{d}\mathbf{m}. \quad (18)$$

The objective function (17) can be maximized by means of any numerical optimization scheme, simple (like gradient descent) or more sophisticated.

#### **Extension to multi-dimensional systems.**

It is important to stress that the proposed scheme is by no means limited to scalar dynamical models. It is true that the analytic expressions for the Fisher metric and the associated geodesics are known for scalar outputs. This allows to derive elegant analytic expressions for the inverse volume (17) to maximize. However, in the case of multi-dimensional linear systems both Fisher metric and its geodesics can still be computed by means of an iterative numerical scheme [28, 27]. The extension of the pullback manifold learning scheme to multi-dimensional systems is then straightforward.

In any case, using the Fisher information metric as basis Riemannian metric is not mandatory. We can as easily adopt the standard Euclidean metric as initial distance, and build a family of pullback Euclidean metrics to optimize upon.

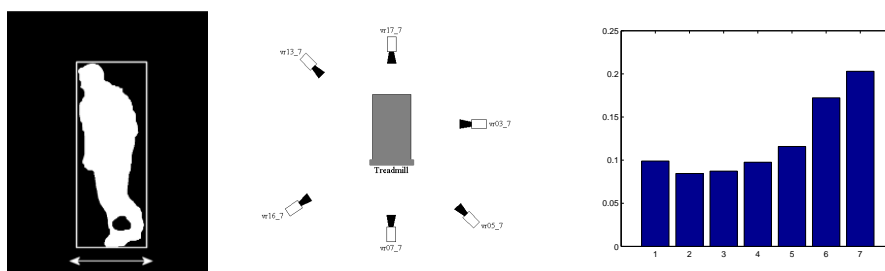
## **5 Effect on classification performances**

It can be argued that maximizing classification performance and minimizing volume elements are indeed correlated optimization problems. In other words, that classification is easier on the pullback manifold which better interpolates the dataset of dynamical systems.

**Experiments on human motion classification.** To test this conjecture we considered the problem of recognizing actions and identities from image sequences. We used the Mobo database [19], 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 fixed viewpoints (see Figure 3-middle). Each sequence of the database possesses three labels: action, view, and identity. As the objective function to optimize has been computed

in Section 4 for single input-single output linear systems, we extracted from each image a *single scalar feature*, the width of the minimum box containing the silhouette (Figure 3-left). Each scalar sequence was passed as input to an identification algorithm which generated a dataset of  $AR(2)$  linear systems, one for each labeled motion sequence. However, as we just pointed out, this is not a limitation of the proposed approach.

We empirically assessed the conjecture on the effect of pullback manifold learning on classification by measuring the performance of a nearest-neighbor classifier based on the optimal pullback metrics induced by the diffeomorphism (10) and comparing the results with those of NN classifiers based on several a-priori distances between models (including the Fisher geodesic distance itself).



**Fig. 3.** Left: from each image the width of the bounding box containing the silhouette is extracted. Middle: 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 the NN classifier in the ID recognition experiment, based on several a-priori distances between dynamical models and two pullback metrics induced by the Fisher geometry. Classification rate for the used metrics averaged on the size of the training set. 1 - Frobenius norm of the matrices in canonical form; 2 - gap metric; 3-  $\nu$ -gap metric; 4 - subspace angles; 5 - basis Fisher metric; 6 - pullback metric with diffeomorphism (19); 7 - pullback metric with diffeomorphism (10).

**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 with (the series of bounding box widths along the image sequence) clearly provides insufficient information. However, measuring the comparative performance of the metrics can be useful to see how learning appositively a metric improves classification. We implemented a naive Frobenius norm of the system matrices in canonical form, gap and  $\nu$ -gap metrics [14, 15], subspace angles [13, 12], and of course the Fisher geodesic distance together with the associated optimal pullback metric.

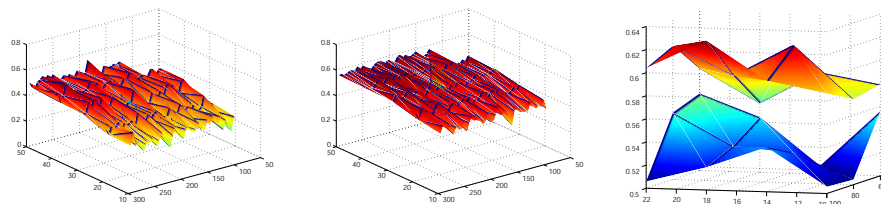
Figure 3-right shows the *average* percentage of correctly classified testing systems

over several dozens of runs in which we randomly selected an increasing number of systems in both training and testing set. The number of training sequences ranged from 10 to 50, while the size of the testing set would vary from 50 to 300. The purpose of the random selection was avoiding biased results due to a particular choice of training and testing sets. Standard variations were very small and they are not reported. As a comparison we also tested a second diffeomorphism for  $AR(2)$  systems, namely

$$F_\lambda(\mathbf{m}) = [\lambda m_1 + (1 - \lambda)m_2, \lambda m_2 + (1 - \lambda)m_3, \lambda m_3 + (1 - \lambda)m_1], \quad (19)$$

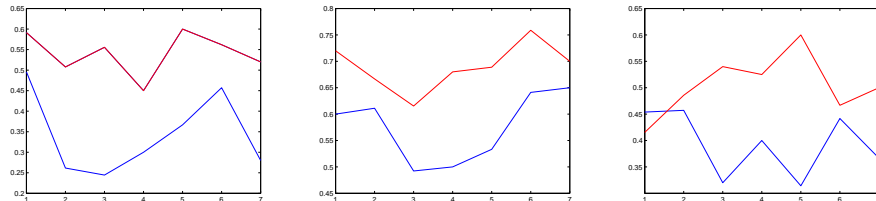
with  $0 < \lambda < 1$ . Figure 3-right shows that the larger the parameter space (as (19) has just one parameter, while (10) has two) the greater the chance of selecting a better metric. All other distances have rather similar performances as they obviously have not been designed to solve classification tasks but for identification purposes. In the following only the performance of the second best distance is reported for comparison.

**Action recognition.** In a second series of experiments we used the pullback geodesic distance to classify the *action* performed by the person. The correct classification rate was measured after several dozens of repeated trials in which training and testing set of increasing size (as in the ID experiment) were randomly selected from the database in order to avoid any bias in the results. Figure



**Fig. 4.** Average recognition performance of second-best a-priori distance (left, in this case the Frobenius distance between the matrices representing two systems) and optimal pullback metric for  $AR(2)$  systems and diffeomorphism (19) (middle) in the action recognition experiment. Here the whole dataset is considered, regardless of the viewpoint. Rates are plotted against the size of testing ( $x$  axis) and training ( $y$  axis) set. Right: recognition rates for view 5 only. The learnt pullback metric always outperforms its best competitor.

4 compares again the average recognition rate over such repetitions of the optimal pullback metric for autoregressive systems (in particular under deformation map (19)) and its best competitor *among a-priori metrics* (typically Frobenius norm or cepstrum distance), as a function of testing and training size, 10 repetitions for each pair of such sizes. Rates are not very high, but remember that actions in the Mobo database are just slightly different variations of the walking gait, and we were using a *single* scalar feature. We want to stress relative



**Fig. 5.** Correct action classification rates for sequences coming from three different viewpoints (1,5,6). Sequences are represented as autoregressive order 2 systems. The performance (in red) of the pullback metric associated with the map (10) is shown outperforming the best competitor (in blue, typically Frobenius norm or Martin’s distance) for increasing sizes of the testing set (from 20 to 50).

performances here. 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.

Figure 5 illustrates the average performance of the classifiers associated with the optimal pullback metric for  $AR(2)$  models with (this time) diffeomorphism (10) and the best competing non-learned, a-priori distance (gap,  $\nu$ -gap, etc.) for four view-dependent experiments, as a function of the testing set size. As usual the average (for each size) is computed over 10 repeated random selections. The optimal pullback metric performs far better than all the others.

## 6 Conclusions

In this paper we proposed a differential-geometric framework for manifold learning given a data-set of linear dynamical models. We pose an 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 showed experimental results concerning identity and action recognition, which support the claim that learning a metric in such a way actually improves classification performance with respect to the Fisher metric and all other known a-priori distances between dynamical models.

## References

1. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning distance functions using equivalence relations. In: ICML03. (2003) 11–18
2. Bilenko, M., Basu, S., Mooney, R.: Integrating constraints and metric learning in semi-supervised clustering. In: Proc. of ICML’04. (2004)
3. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. In: Advances in Neural Information Processing Systems. (2004)

4. Tsang, I., Kwok, J., Bay, C., Kong, H.: Distance metric learning with kernels. In: Proceedings of the International Conference on Artificial Intelligence. (2003)
5. Zhang, Z.: Learning metrics via discriminant kernels and multidimensional scaling: Toward expected euclidean representation. In: ICML'03. (Hong Kong, 2003)
6. 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)
7. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290(5500)** (2000) 2323–2326
8. Bengio, Y., Paiement, J.F., Vincent, P.: Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering. Technical report, Université de Montreal (2003)
9. 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)
10. Shental, N., Hertz, T., Weinshall, D., Pavel, M.: Adjustment learning and relevant component analysis. In: ECCV'02. (2002)
11. Doretto, G., Soatto, S.: Editable dynamic textures (2002)
12. Martin, R.J.: A metric for ARMA processes. *IEEE Trans. on Signal Processing* **48(4)** (April 2000) 1164–1170
13. Cock, K.D., Moor, B.D.: Subspace angles and distances between arma models. *Systems and Control Letters* (2002)
14. 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
15. Vinnicombe, G.: A  $\nu$ -gap distance for uncertain and nonlinear systems. In: Proc. of CDC'99. (Phoenix, 1999)
16. Vidyasagar, M.: The graph metric for unstable plants and robustness estimates for feedback stability. *AC* **29** (1984) 403–417
17. Smola, A., Vishwanathan, S.: Hilbert space embeddings in dynamical systems. In: Proc. of IFAC'03. (August 2003) 760 – 767
18. Kullback, S., Leibler, R.A.: On information and sufficiency. *Ann. Math. Stat.* **22** (1951) 79–86
19. Gross, R., Shi, J.: The CMU motion of body (Mobo) database. Technical report, CMU (2001)
20. Hanzon, B.: Identifiability, recursive identification and spaces of linear dynamical systems. *CWI Tracts* **63-64** (Amsterdam 1989)
21. Ravishanker, N., Melnick, E., Tsai, C.L.: Differential geometry of ARMA models. *J. Time Series Anal.* **11** (1990) 259–274
22. Rijkeboer, A.: Fisher optimal approximation of an AR(n)-process by an AR(n-1)-process. In: Proceedings of ECC'93. (1993) 1225–1229
23. Lebanon, G.: Metric learning for text documents. *PAMI* **28(4)** (2006) 497–508
24. Amari, S.I.: *Differential geometric methods in statistics*. Springer-Verlag (1985)
25. Murray, M., Rice, J.: *Differential Geometry and Statistics*. CRC Press (1993)
26. Belkin, M., Niyogi, P.: Semi-supervised learning on riemannian manifolds. *Machine Learning* **56** (2004) 209–239
27. Hanzon, B., Peeters, R.: Aspects of Fisher geometry for stochastic linear systems, problem 25. In: *Open Problems in Mathematical Systems and Control Theory*. (2002) 27–30
28. Peeters, R., Hanzon, B.: On the Riemannian manifold structure of classes of linear systems. In: *Equadiff2003*. (2003)