On the Distance Between Non-stationary Time Series

Stefano Soatto

Computer Science Department
University of California, Los Angeles
soatto@ucla.edu

1 Introduction

Comparing time series is a problem of critical importance in a broad range of applications, from data mining (searching for temporal "patterns" in historical data), to speech recognition (classifying phonemes from acoustic recordings), surveillance (detecting unusual events from video and other sensory input), computer animation (concatenating and interpolating motion capture sequences), just to mention a few. The problem is difficult because the same event can manifest itself in a variety of ways, with the data subject to a large degree of variability due to *nuisance factors* in the data formation process. For instance, the presence of a person walking in a video sequence can vary based on the individual, his gait, location, orientation, speed, clothing, illumination etc. And yet, if I see Giorgio Picci, I can recognize him from one hundred yards away by the way he walks, regardless of what he is wearing, or whether it is a sunny or a cloudy day. One could conjecture that there must exist some statistics of my retinal signals that are invariant, or at least insensitive, to such nuisance factors and are instead Giorgio-Specific (GS). The information ought to be encoded in the temporal evolution of the retinal signals, for one can strip the images of their pictorial content by attaching light bulbs to one's joints and turning off the lights (or use a state-of-the-art motion capture system, for instance one made by E-motion/BTS); one can still tell a great deal from just the moving dots [6].

To be sure, searching for invariant GS statistics is not the only way to get rid of the nuisances: One can also eliminate them as part of the matching process when comparing two time series. Let us consider, for example, the nuisance of the initial observation instant, t_0. If we observe the temporal evolution of joint positions and think of them as trajectories $\{y_1(t)\}_{t\in\mathbb{R}}$ in, say, $\mathbb{L}^2$, we could compare it to a sample sequence from a database, $\{y_2(t)\}_{t\in\mathbb{R}}$, by sliding one on top of the other until the $\mathbb{L}^2$ norm of the difference is minimized: $d_0(y_1, y_2) = \min_{t_0} \int \|y_1(t) - y_2(t-t_0)\|^2 dt$. A more elegant and efficient solution is to seek for a statistic of each time series, i.e. a deterministic function $\Sigma_i \doteq \phi(y_i)$, that is invariant with respect to t_0, and then to compare such statistics directly. For the case of sequences that admit statistics that are invariant with respect to t_0, a.k.a. *stationary*, this can be done, and the resulting *realization theory* is a success story of Systems Theory, one where Giorgio Picci and his collaborators have played a key role [11]. Endowing the space of realizations with a metric and a probabilistic

A. Chiuso et al. (Eds.): Modeling, Estimation and Control, LNCIS 364, pp. 285–299, 2007.
springerlink.com

structure is the key to enabling a new level of finesse in the applications cited above, where we are not simply trying to detect the presence of a person and whether he is walking, but also to tell gender, identity, even state of mind. Some preliminary steps in this direction have shown promise in the recognition of stationary human motion [2].

In this work, I wish to *extend the formulation of the problem to more general classes of nuisances* for time series that are not stationary. In the spirit of realization theory, I will forgo the prevailing approach where sequences are compared via their likelihood. In this approach, the similarity between two sequences is measured by how well the model of one (say a realization $\Sigma_1 = \phi(y_1)$) "explains" the other, for instance quantified by the covariance of the innovation process of y_2. In this approach, the more data are available, the better the estimate of Σ_1, the worse the classification error is, an apparent paradox induced by the fact that the generalization model underlying this approach is trivial: Each realization models one sequence and noisy versions of it, without regard for the structure of the intrinsic variability that different realizations of the same process exhibit. I will also not go as far as invoking the full power of chaotic non-linear models as customary in the physics literature [7], because the processes of interest are usually observed on short time-scales and do not exhibit chaotic behavior. Instead, I will attempt to define distances between processes that are invariant, or at least insensitive, to specific classes of nuisances. Once these are available, one can attempt to construct probability distributions and therefore define priors in the space of time series, compute likelihoods, perform optimal decisions etc. in the way one would do if nuisances were absent. For instance, in $\mathbb{L}^2$ one can define a distance

$$d_0(y_1, y_2) = \int_0^T \|y_1(t) - y_2(t)\|^2 dt \tag{1}$$

and from this, with some caveats, define a probabilistic structure $dP(y)$. In our case, we can think of events of interest as equivalence classes under the action of a nuisance group, and therefore we need to define a suitable quotient space (or base) to perform the comparison. Although ideally one would want a true Riemannian metric (homogeneous spaces are in general not flat) and integrate it along geodesics to compute distances, we will limit ourselves to defining cord distances directly. We will do so in steps, first introducing nuisances in general, then describing a variety of distances. For simplicity we will assume that sequences are observed over a common finite interval $[0,\ T]$, although most of the considerations can be extended to the case where the initial time and the duration are also included among the nuisances.

2 Formalization

The first step to introduce nuisances is to re-write (1) in a slightly different way as

$$d_0(y_1, y_2) = \min_h \int_0^T \|y_1(t) - h(t)\|^2 + \|y_2(t) - h(t)\|^2 dt. \tag{2}$$

Note that the two expressions (1) and (2) are identical up to a factor of two, as the right hand-side of (2) is bounded from below by (1), and from above by the same quantity by choosing $\hat{h} = (y_1 + y_2)/2$. Note also that we have been deliberately vague as to the space where h lives, which we will indicate by $\mathcal{H}$; despite it being infinite-dimensional, we do not need to impose regularization on h to solve the optimization above, which is trivially done in closed form. The reason for introducing such an auxiliary variable $h \in \mathcal{H}$ will become clear shortly, but already one can see that this writing highlights the underlying data-formation model: Both time series are generated from some (deterministic but) unknown function h, corrupted by two different realizations of additive "noise" (here the word noise lumps all unmodeled phenomena, not necessarily associated to sensor errors)

$$y_i(t) = h(t) + n_i(t) \quad i = 1, 2;\ t \in [0,\ T] \tag{3}$$

where, for instance, $n_i(t) \overset{iid}{\sim} \mathcal{N}(0, \Sigma)\ \forall\ t \in [0, T]; i = 1, 2$. Under this model, the distance is obtained by finding the (maximum-likelihood) solution for h that minimizes

$$\phi_{data}(y_1, y_2|h) \doteq \sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt \tag{4}$$

subject to (3). Note that an obvious interpretation of h is that of the *average* of the two time series. This will become handy later. Although not necessary at this stage, one could consider regularized distances, for instance

$$d_{reg}(y_1, y_2) = \min_{h \in \mathcal{H}} \phi_{data}(y_1, y_2|h) + \phi_{reg}(h) \tag{5}$$

where, for instance, $\phi_{reg}(h) = \int_0^T \|\nabla h\| dt$. Once a distance is available, one can perform classification in a number of ways, for instance using simple k-nearest neighbors [2].

2.1 Introducing Nuisances

Let us now consider some simple nuisances of the data collection process, and how to eliminate them in computing a meaningful notion of distance between time series. We have already discussed the role of the initial condition $t_0 = \beta$, corresponding to a model $y(t) = h(t + \beta) + n(t)$. Another common accident of data collection is a different sampling frequency, which translates into an affine deformation of the temporal axis $y(t) = h(\alpha t + \beta) + n(t)$. A slightly more elaborate model is a projective transformation of the temporal axis $y(t) = h(\frac{\alpha t + \beta}{\gamma t + \delta}) + n(t)$. In order to generalize this model, we have to resort to infinite-dimensional groups of domain diffeomorphisms of the interval $[0, T]$ [14]; the data formation model (3) above then becomes

$$y_i(t) = h(x_i(t)) + n_i(t) \quad i = 1, 2. \tag{6}$$

Correspondingly, the data term of the cost functional we wish to optimize is

$$\phi_{data}(y_1, y_2|h, x_1, x_2) \doteq \sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt \tag{7}$$

from which it is clear that the model is over-determined, and we must therefore impose regularization [9] in order to compute

$$d_1(y_1, y_2) = \min_{h \in \mathcal{H}, x_i \in \mathcal{U}} \phi_{data}(y_1, y_2 | h, x_1, x_2) + \phi_{reg}(h). \tag{8}$$

The functions $x_i \in \mathcal{U}$ are called *time warpings*, and in order for $\tau \doteq x(t)$ to be a viable temporal index, x must satisfy a number of properties. The first is continuity (time, alas, does not jump); in fact, it is common to assume a certain degree of smoothness, and for the sake of simplicity we will assume that x_i is infinitely differerentiable. The second is causality: The ordering of time instants has to be preserved by the time warping, which can be formalized by imposing that x_i be monotonic. Additional constraints can be imposed that either are specific to a particular application, or to make the mathematical treatment simpler. A common choice is to impose $x_i(0) = 0; x_i(T) = T$ so that the interval $[0, T]$ is fixed by the warping function. These regularization constraints on x_i are implicit in the notation $x_i \in \mathcal{U}$ and will be made explicit in Sect. 3. In the absence of additional constraints, the solution of this problem leads to a well-known technique which we describe next.

2.2 Dynamic Time Warping

The solution of (8) (or the determination of the minimizers $\hat{h}, \hat{x}_i$) is called *dynamic time warping* (DTW) and is standard practice in speech processing as well as in temporal data mining. Making the constraints more explicit, we can re-write the distance above as

$$\boxed{d_2(y_1, y_2) = \min_{h \in \mathcal{H}, x_i \in \mathcal{U}} \sum_{i=1}^{2} \int_0^T \|y_i(t) - h(x_i(t))\|^2 + \lambda \|\nabla h(t)\| dt} \tag{9}$$

where λ is a tuning parameter for the regularizer that is not strictly necessary for this model, and can be set equal to zero, for instance, by choosing $h(t) = y_1(x_1^{-1}(t))$ (or, similarly to what we have done earlier $h(t) = (y_1(x_1^{-1}(t)) + y_2(x_2^{-1}(t))/2)$ and $x(t) \doteq x_2(x_1^{-1}(t))$. This yields a "reduced" optimization problem with only one (functional) variable,

$$\int_0^T \|y_1(t) - y_2(x(t))\| dt + \mu \overline{\phi}_{reg}(x) \tag{10}$$

where we have assumed that $\mathcal{U}$ can be defined algebraically as $\{x \mid \overline{\phi}_{reg}(x) = 0\} \doteq \mathcal{U}$, as we will make explicit in Sect. 3. In this simplified model, x matches data-to-data, rather than each x_i matching each data y_i to a common underlying template h. This distance relates to the Skorohod topology introduced for time-of-arrival processes to account for small temporal jittering [1]. The reason for solving a seemingly more complex problem (9), rather than (10), is because, in the presence of noise in the measurements y_1, y_2, the warping x in (10) will attempt to fit the noise, causing the minimizer $\hat{x}$ to be highly irregular. This is usually addressed by enforcing heavy regularization (large $\mu \in \mathbb{R}_+$.) The advantage of the auxiliary variables x_i, h, as discussed in detail in [17], is to avoid warping the (noisy) data y_i, but instead to warp the (smooth) template h; because in general DTW is non-linear and one has to resort to gradient algorithms based on the first-order (Euler-Lagrange) optimality conditions, the presence of an explicit model h allows one to "push" the derivatives onto the model, arriving at *gradient-based* algorithms that *do not involve differentiation of the (noisy) data* [22].

Before we elucidate the structure of the space of warping functions $\mathcal{U}$, we pause to note that in general it is an infinte-dimensional group of diffeomorphisms [4]), as it is (at least locally) invertible. Under the action of this group we can now distinguish two scenarios:

- $\mathcal{U}$ acts transitively on $\mathcal{H}$: For any given y_1, y_2, there exists at least a $\hat{h} \in \mathcal{H}$ (a "template" in Grenander's nomenclature) and $\hat{x}_1, \hat{x}_2 \in \mathcal{U}$ such that the data term $\phi_{data}(y_1, y_2|\hat{h}, \hat{x}_1, \hat{x}_2)$ is identically zero. In other words, with a group action x one can reach any y from some h. In this case, the data term of the distance is zero, and the actual distance reflects the amount of "energy" or "work" necessary to reach it. This is quantified by the regularization terms $\overline{\phi}_{reg}(x)$.
- $\mathcal{U}$ is restricted (for instance, it belongs to a parametric class of functions), in which case it can only bring h "close" to y, and their proximity is reflected in the distance.

In either case, the minimizer $\hat{h}$ can be interpreted as the "average" of the data, in the sense elucidated in [16]. As we have mentioned, the equivalence classes $[y] \doteq \{y_i \circ x_i^{-1} \,|x_i \in \mathcal{U}\}$ represent the objects of interest, and the quotient space can be thought of as the space where comparison is to be performed [20].

2.3 Dynamics, or Lack Thereof, in DTW

It is important to note that *there is nothing "dynamic" about dynamic time warping.*[1] There is no requirement that the warping function x be subject to physical constraints, such as the action of forces, the effects of inertia etc. However, some notion of dynamics can be coerced into the problem by characterizing the set $\mathcal{U}$ in terms of the solution of a differential equation. Following [15], as shown by [12], one can represent allowable $x \in \mathcal{U}$ in terms of a small, but otherwise unconstrained, scalar function u: $\mathcal{U} = \{x \in \mathcal{H}^2([0,\ T]) \,|\ddot{x} = u\dot{x};\ u \in \mathbb{L}^2([0,T])\}$ where $\mathcal{H}^2$ denotes a Sobolev space. If we define $\rho_i \doteq \dot{x}_i$ then $\dot{\rho} = u\rho$; we can then stack the two into $\xi \doteq [x,\ \rho]^T$, and $C = [1,\ 0]$, and write the data generation model as

$$\begin{cases} \dot{\xi}_i(t) = f(\xi_i(t)) + g(\xi_i(t))u_i(t) \\ y_i(t) = h(C\xi_i(t)) + n_i(t) \end{cases} \tag{11}$$

as done by [12], where $u_i \in \mathbb{L}^2([0,T])$. Here f, g and C are given, and $h, x_i(0), u_i$ are nuisance parameters that are eliminated by minimization of the data term

$$\phi_{data}(y_1, y_2|h, x_1(0), x_2(0), u_1, u_2) \doteq \sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt \tag{12}$$

[1] The name comes from the fact that a discretized version of this problem can be solved using dynamic programming, since the integral in (8) can be decomposed into a sum of cost-to-go terms due to the monotonicity constraint.

subject to (11), with the addition of a regularizer $\lambda\phi_{reg}(h)$ and an energy cost for u_i, for instance $\phi_{energy}(u_i) \doteq \int_0^T \|u_i\|^2 dt$. Writing explicitly all the terms, the problem of dynamic time warping can be written as

$$\boxed{d_3(y_1, y_2) = \min_{h\in\mathcal{H}, u_i\in\mathbb{L}^2, x_i(0)} \sum_{i=1}^{2} \int_0^T \|y_i(t) - h(C\xi_i(t))\| + \lambda\|\nabla h(t)\| + \mu\|u_i(t)\| dt} \tag{13}$$

subject to $\dot{\xi}_i = f(\xi_i) + g(\xi_i)u_i$. Note, however, that this differential equation does not arise out of the desire to enforce dynamic constraints exhibited in the data, but it is only an expedient to (softly) enforce causality by imposing a small "time curvature" u_i.

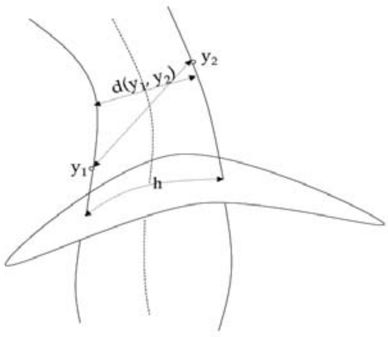

Fig. 1. Time series are points along equivalence classes (fibers), where the nuisance acts as a group that moves points along their fibers. A proper distance in the homogeneous space of equivalence classes is independent of the position of points on the fibers: This can be achieved by minimization, for instance by finding the minimum distance among the fibers (dotted line), or by canonization, by finding a base of the fiber bundle where comparisons are made (dashed lines). The distance proposed in [13] only moves one of the two points (or their average) along the fiber. One of the byproducts of the computation of the distance is the average between the data, a concept that extends to any number $N \geq 2$ time series.

Note also that the solution of the minimization above is not tantamount to a nonlinear system identification task. In fact, in system identification one is given *one* time series y, with the task of inferring the model parameters h, possibly along with the state, input and initial condition (here f, g and C are given). If that were the case, we could always choose $h = y$ for any state, input and initial condition, making the problem trivial. Here instead we are given *two* time series, and we want to jointly estimate the unknown parameters of a model h that, under suitable inputs u_i, can generate the data with minimal discrepancy, measured by ϕ_{data}.

3 Time Warping Under Dynamic Constraints

In this section we introduce a notion of time warping that respects the dynamic structure of the data. Some preliminary progress towards this goal has been made by [8], who

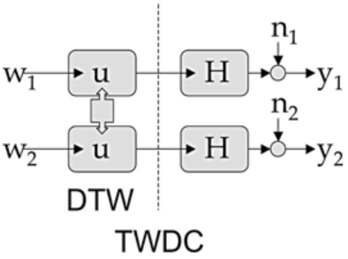

Fig. 2. Traditional dynamic time warping (DTW) assumes that the data come from a common function that is warped in different ways to yield different time series. In time warping under dynamic constraints (TWDC), the assumption is that the data are output of a dynamic model, whose inputs are warped versions of a common input function.

extended the warping function to include derivatives. However, no specific model or assumption, other than small velocity difference, is made between the two time series. Instead, we look for warpings that are compatible with the dynamics imposed by the forces and inertias of the physical processes that generated the data y_i. We will address this issue by considering a generalization of the model (11). The basic idea is illustrated in Figure 2: Rather than the data being warped versions of some common function, as in (6), we will assume that *the data are outputs of dynamical models driven by inputs that are warped versions of some common function.* In other words, given two time series $y_i,\ i = 1, 2$, we will assume that there exist suitable matrices A, B, C, state functions x_i of suitable dimensions, with their initial conditions, and a *common input* u such that the data are generated by the following model, *for some warping functions* $w_i \in \mathcal{U}$*:*

$$\begin{cases} \dot{x}_i(t) = Ax_i(t) + Bu(w_i(t)) \\ y_i(t) = Cx_i(t) + n_i(t). \end{cases} \tag{14}$$

Our goal is to find the distance between the time series by minimizing with respect to the nuisance parameters the following data discrepancy:

$$\phi_{data}(y_1, y_2|u, w_i, x_i(0)) \doteq \sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt \tag{15}$$

subject to (14), together with regularizing terms $\overline{\phi}_{reg}(u)$ and with $w_i \in \mathcal{U}$. Notice that this model is considerably different from the previous one, as the state ξ earlier was used to model the temporal warping, whereas now it is used to model the data, and the warping occurs at the level of the input. It is also easy to see that the model (14), despite being linear in the state, includes (11) as a special case, because we can still model the warping functions w_i using the differential equation in (11). In order to write this *time warping under dynamic constraint* problem more explicitly, we will use the following notation:

$$y(t) = Ce^{At}x(0) + \int_0^T Ce^{A(t-\tau)}Bu(w(\tau))d\tau \doteq L_0(x(0)) + L_t(u(w)) \tag{16}$$

in particular, notice that L_t is a convolution operator, $L_t(u) = F * u$ where F is the transfer function. We first address the problem where A, B, C (and therefore L_t) are given. For simplicity we will neglect the initial condition, although it is easy to take it into account if so desired. In this case, we define the distance between the two time series

$$\boxed{d_4(y_1, y_2) = \min \sum_{i=1}^{2} \int_0^T \|y_i(t) - L_t(u_i(t))\| + \lambda \|u_i(t) - u_0(w_i(t))\| dt} \tag{17}$$

subject to $u_0 \in \mathcal{H}$ and $w_i \in \mathcal{U}$. Note that we have introduced an auxiliary variable u_0, which implies a possible discrepancy between the actual input and the warped version of the common template. This problem can be solved in two steps: A deconvolution, where u_i are chosen to minimize the first term, and a standard dynamic time warping, where w_i and u_0 are chosen to minimize the second term. Naturally the two can be solved simultaneously.

3.1 Going Blind

When the model parameters A, B, C are common to the two models, but otherwise unknown, minimization of the first term corresponds to blind system identification, which in general is ill-posed barring some assumption on the class of inputs u_i. These can be imposed in the form of generic regularizers, as common in the literature of blind deconvolution [3]. This is a general and broad problem, but beyond our scope here, so we will forgo it in favor of an approach where the input is treated as the output of an auxiliary dynamical model, also known as *exo-system* [5]. This combines standard DTW, where the monotonicity constraint is expressed in terms of a double integrator, with TWDC, where the actual stationary component of the temporal dynamics is estimated as part of the inference. The generic warping w, the output of the exo-system (see Figure 3), satisfies

$$\begin{cases} \dot{w}_i(t) = \rho_i(t), \quad i = 1, 2 \\ \dot{\rho}_i(t) = v_i(t)\rho_i(t) \end{cases} \tag{18}$$

and $w_i(0) = 0, \ w_i(T) = T$. This is a multiplicative double integrator; one could conceivably add layers of random walks, by representing v_i are Brownian motion. Combining this with the time-invariant component of the realization yields the generative model for the time series y_i:

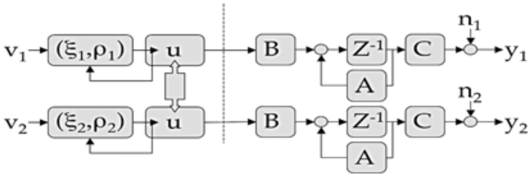

Fig. 3. TWDC can be modeled as comparison of the output of two dynamical models driven by two exo-systems that are in charge of time-warping a common input u

$$\begin{cases} \dot{w}_i(t) = \rho_i(t), \quad i = 1, 2 \\ \dot{\rho}_i(t) = v_i(t)\rho_i(t) \\ \dot{x}_i(t) = Ax_i(t) + Bu(w_i(t)) \\ y_i(t) = Cx_i(t) + n_i(t). \end{cases} \tag{19}$$

Note that the actual input function u, as well as the model parameters A, B, C, are common to the two time series. A slightly relaxed model, following the previous subsection, consists of defining $u_i(t) \doteq u(w_i(t))$, and allowing some slack between the two; correspondingly, to compute the distance one would have to minimize the data term

$$\phi_{data}(y_1, y_2 | u, w_i, A, B, C) \doteq \sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt \tag{20}$$

subject to (19), in addition to the regularizers

$$\overline{\phi}_{reg}(v_i, u) = \sum_{i=1}^{2} \int_0^T \|v_i(t)\|^2 + \|\nabla u(t)\|^2 dt \tag{21}$$

which yields a combined optimization problem

$$\boxed{d_5(y_1, y_2) = \min_{u, \in \mathbb{L}^2, A, B, C} \sum_{i=1}^{2} \int_0^T (\|y_i(t) - Cx_i(t)\|^2 + \|v_i(t)\|^2 + \|\nabla u(t)\|^2) dt} \tag{22}$$

subject to (19).

3.2 Computing the Distance

In order to compute the distance for the various cases defined above we have to solve what is in effect an optimal control problem. Specifically, for the following model[2]

$$\begin{cases} \dot{x} = v \quad x(0) = 0; \; x(T) = T \\ \dot{v} = uv \end{cases} \tag{23}$$

relative to the cost function

$$J = \int_0^T \|y - h(x)\|^2 + u^2 dt \tag{24}$$

we are looking for $\min J$ subject to (23). The Hamiltonian is given by

$$H(x, u, \lambda) = \|y - h(x)\|^2 + u^2 + \lambda^T [v, uv]^T \tag{25}$$

and the value function (optimal cost) V should satisfy the Hamilton-Jacobi-Bellmann equation

[2] We use the simplified version where $y_1 = y, \; y_2 = h$, but all considerations can be easily extended to the more general case, where u_i, v_i, x_i have $i = 1, 2$.

$$\frac{\partial V}{\partial t}(x,t) + \min_u H\left(x, u, \frac{\partial V}{\partial x}\right) = 0 \tag{26}$$

with suitable boundary conditions. A discretized version of this equation can be computed on a sample time sequence

$$V(x,t) = \min_u \int_t^{t+\Delta} \|y - h(x)\|^2 + u^2 d\tau + V(x(t+\Delta), t+\Delta) \tag{27}$$

with boundary conditions for x to satisfy (23) using Dynamic Programming. If we are content with a (faster) local gradient algorithm based on the first-order (Euler-Lagrange) optimality conditions, then we simply update iteratively u, starting from an initial estimate, in the direction opposite to the gradient of the Hamiltonian.

In the case of TWDC, the additional (finite-dimensional) unknowns due to the model parameters (A, B, C) can be easily incorporated. Note that the quotient structure of the parameter space of realizations is not an issue here since all that matters is the minimum value (distance), rather than the minimizer (realization).

4 Correlation Kernels for Non-stationary Time Series

In this section we explore a distinctly different approach to defining a distance between time series. Instead of computing the data term ϕ_{data} in terms of the $\mathbb{L}^2$ distance between the time series, an alternative consists in defining an inner product between the two, via correlation, from which a cord distance can be easily computed. This has been done for the case of time-invariant models in [21]. In this section we illustrate how these concepts can be generalized to allow of time-warpings of the input. Using the notation in (16), we define the (symmetric, positive-definite) kernel

$$K(y_1, y_2|u) \doteq \mathbb{E}_{v_1,v_2}\left[\text{trace} \int_0^T y_1(t) y_2^T(t) d\mu(t)\right]$$

$$= \mathbb{E}_{v_1,v_2}\left[\text{trace} \int_0^T L_t(u(w_1(t))) L_t^T(u(w_2(t))) d\mu(t)\right] \tag{28}$$

where $d\mu(t) \sim \frac{e^{-\lambda t}}{t}$ includes an exponential discounting term and the expectation is computed with respect to the joint density of v_1, v_2, subject to (18). We can make the functional explicit by exploiting the calculations leading to equation (6.11) of [15], to obtain

$$L_t = C \int_0^T e^{A(t-\tau)} Bu \left(K_0 + K_1 \int_0^\tau \exp \int_0^{\tau'} v(\tau'') d\tau'' d\tau'\right) d\tau. \tag{29}$$

This can be substituted into the previous equation and integrated against the joint density of v_1 and v_2. Several simplifying assumptions are possible for this density: One can assume, as in [21], that the two are independent, or that they are identical. One could also assume that v_1, v_2 are small an independent, an assumption implicit in the choice of regularizer in (22) (the second term in the integral). This can be enforced in practice by choosing a joint density for discretized versions of v_1, v_2 proportional to $\exp(-(\|v_1\|^2 + \|v_2\|^2))$. The two constants K_0, K_1 can be set by imposing the boundary conditions. Note that the kernel depends upon the model parameters $\{A, B, C\}$, hidden in the operator L_t, as well as on the unknown input $u \in \mathcal{U}$. Note also that the initial condition can be used to define an additive kernel, identically to what done in [21]. The kernel above satisfies Mercer's condition, and because the sum of Mercer kernels is also Mercer, this procedure yields a viable kernel in a straightforward manner.

The non-straightforward part of this program is the computation of the expectation above, for which no better strategy than general Monte Carlo is currently available. However, assuming that it can be done, one can use the kernel to define a distance via

$$\phi_{data}(y_1, y_2|u, A, B, C) \doteq K(y_1, y_1|u) + K(y_2, y_2|u) - 2K(y_1, y_2|u) \qquad (30)$$

and then optimize with respect to the unknowns $u \in \mathcal{U}, A, B, C$. As an alternative, one could marginalize the unknowns to compute

$$\boxed{d_6(y_1, y_2) = \int (\phi_{data} + \lambda \phi_{reg}(u)) dP(u)} \qquad (31)$$

as an alternative to extremization when a measure on $\mathcal{U}$ is available.

5 Invariance Via Canonization

In previous sections we have explored various alternative distances where the nuisances were eliminated by extremization (i.e. solving an optimization, or "search," problem), or by marginalization. In either cases, the computation of the distance entails the solution of a difficult computational problem. As we have pointed out at the beginning, an alternative way to endow a homogenous space with a metric structure is to reduce it to its base, that is to define for each class a *canonical representative*, and then to compute a distance in the base space that respects its geometry. For the case of stationary processes, a variety of canonical realizations has been defined. In our case, the canonical representative would have to be a diffeomorphism of the domain $[0, T]$, and "canonical" refers to the fact that given a certain time series $\{y(t)\}_{t \in [0,T]}$ and its associated equivalence class $[y] = \{y(w(t)), w \in \mathcal{U}\}$, a canonical representative $\hat{y} \doteq y(\hat{w})$ must be computed solely from $[y]$, i.e. without resorting to comparison with other fibers.

For simplicity, we illustrate the canonization process for the simple case of affine domain deformations first, although the construction can be extended to arbitrary diffeomorphisms as shown in [18] (see also Figure 4 for an illustration on this procedure). For domain transformations of the form $w(t) = \alpha t + \beta$ we have to relax the fixed boundary conditions $w(0) = 0$ (lest $\beta = 0$) and $w(T) = T$ (lest $\alpha = 1$). Assuming the boundaries of observation of the two sequences to be undetermined, we can canonize each sequence by choosing α and β that generate statistics of $\{y(w(t))\}$ that have a prescribed value. For instance, we can take some differential statistics, of the form $\phi(y) = \frac{d^k y}{dt^k}$, and impose that they take a prescribed value in uniquely identifiable positions. For instance, t_1 can be the first position where $\frac{dy}{dt} = 0$, assigned

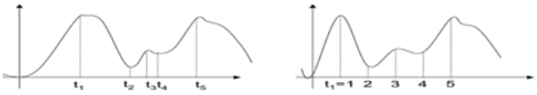

Fig. 4. Simple canonization for comparing time series: Extrema of the warped function (or a scale-space of it) are assigned to fixed value in increasing order. The back-warped time series is then, by construction, invariant to domain warpings.

to a fixed value on the axis, say $t = 0$. This fixes the translation group β. We can then take the second point where $\frac{dy}{dt} = 0$, and assign it to a fixed value, say $t = 1$. This fixes the linear scaling group α. The procedure can be extended to more general warpings, for different statistics, including integral ones (moments) of higher order. Note that if the data y are such that such distinct points do not exist (for instance if $y(t)$ is constant), then *any* α and β would do. So, in other words, where we can find statistics that depend on w, we can fix them to canonize w; where the statistics do not depend on w, these are already, by definition, invariant!

We now extend this construction more formally. Consider a set of data (time series) $\{y_i \in M\}_{i=1,\dots,n}$ undergoing the action of a *nuisance group* $w_i \in \mathcal{U}$, to yield their *irked* versions $\tilde{y}_i = w_i y_i \doteq y_i(w_i(t))$. The ambient space, where the irked data live, is the set of orbits $N \doteq M^{\mathcal{U}}$. Now, suppose that there exists a "feature," i.e. a function $\phi : N \to \mathbb{R}^l$, where l is the dimension of $\mathcal{U}$, such that $\phi(w^{-1}\tilde{y}) = 0$ uniquely determines u up to a set of measure zero $K \subset N$.[3] Then, we can eliminate the effect of the nuisance by pre-processing each irked datum via $\hat{y} \doteq \hat{w}^{-1}\tilde{y} \mid \phi(\hat{w}^{-1}\tilde{y}) = 0$ to obtain a *canonical* element $\hat{y} \in M$. The choice of $\hat{y}$ is canonical in the sense of conforming to the rule $\phi(\hat{y}) = 0$.[4] The function ϕ is called a *pontifical feature* since it is the feature (i.e. the statistic) that determines how $\tilde{y}$ is to be canonized. If the data space M, undergoing the action of the nuisance group $\mathcal{U}$, admits a pontifical feature, it is called *sanctifiable*. We write the canonization process more succinctly as

$$\hat{y} \doteq \phi^{-1}(0|\tilde{y}) = \hat{w}^{-1}\tilde{y} \mid \phi(\hat{w}^{-1}\tilde{y}) = 0 \tag{32}$$

or, with an unholy abuse of notation, as $\hat{y} = \phi^{-1}(\tilde{y})$. It is easy to construct examples that show that not all spaces are sanctifiable [18]. Note, however, that whether a space is sanctifiable depends on the base M as well as on the nuisance $\mathcal{U}$. In general, the larger the nuisance, the more difficult it is for the space to be sanctified. Sometimes it is possible for the space to be sanctifiable, but with only one canonical element. That is, the quotient is zero-dimensional, and the entire irked population $\{\tilde{y}_i\}$ is equal under the law $\phi = 0$. In this case we say that the space N *collapses under the nuisance* $\mathcal{U}$.

[3] Such a set of measure zero is the set of data that is invariant under a subgroup $H \subseteq \mathcal{U}$, i.e. the *symmetry* set of H: $K \doteq \{\tilde{y} \mid \phi(w^{-1}\tilde{y}) = \phi(\tilde{y}) \; \forall \, u \in \mathcal{U}\}$.

[4] The choice of zero in the rule $\phi = 0$ is arbitrary, and any other value $\phi = k \neq 0$ would work just as well, yielding a different set of canonical representatives $\hat{y} \in N/\mathcal{U}$. Therefore, in general the base space $N/\mathcal{U}$ where the canonical representatives live is not necessarily equal to M, but it is related to it by a parallel translation $\tilde{w} \in \mathcal{U}$, so that the "true" space is given by $M = \tilde{w}N/\mathcal{U}$. Since any value $k \neq 0$ can be incorporated into the definition of ϕ, the choice of the zero level set in (32) is without loss of generality.

Example 1. In Grenander's "Deformable templates" [4] the objects of interest ("target shapes") are obtained from a common generator (the "template") under the action of an infinite-dimensional group. Because of the assumption that the group acts transitively, the entire world of objects of interest is equivalent under the action, and therefore the space collapses under the nuisance.

To construct a simple pontifical feature, consider simply the function $\phi(\tilde{y}) \doteq \frac{d\tilde{y}}{dt}$, and let $t_1, \ldots, t_N$ be the N local extrema of $\tilde{y}$:

$$t_i \doteq t \mid \phi(\tilde{y}) = 0, \; i = 1, \ldots, N \tag{33}$$

Because $\frac{d\tilde{y}}{dt} = \frac{dy}{dt}\frac{dw}{dt}$ and by assumption $w \in \mathcal{U}$ we have that $\frac{dw}{dt} > 0, \; t \in [0, T]$, we have that the values $\tilde{y}(t_i)$ are independent of w. We can then choose a canonical representative for w by imposing

$$\hat{w}(t_i) = \frac{i}{N+1}T \tag{34}$$

where we have assumed $t_0 = 0$ and $t_{N+1} = T$. The canonical representative of $\tilde{y}$ is then simply given by

$$\hat{y}(t) = \tilde{y}(\hat{w}^{-1}(t)), \quad t \in [0, T]. \tag{35}$$

Finally, the distance between canonical elements can be simply computed in $\mathbb{L}^2$:

$$\boxed{d_7(y_1, y_2) = \int_0^T \|y_1(\hat{w}_1^{-1}(t)) - y_2(\hat{w}_2^{-1}(t))\|^2 dt = \int_0^T \|\hat{y}_1(t) - \hat{y}_2(t)\|^2 dt} \tag{36}$$

Note that this solution is the canonization counterpart of DTW presented in Sect. 2.2. In order to extend this to TWDC as presented in Sect. 3 one would have to canonize $v(t)$, rather than $w(t)$, which can be done at the expense of additional notation, and we will therefore forgo it in this venue.

As we have already observed, the choice of canonical element is arbitrary, so that the effects of canonization in classification largely depend on the fine art of choosing a pontifical feature. Such a feature is, by design, invariant to the nuisances we have modeled explicitly: The art consists in making it also robust, or "insensitive," to other factors that we do not have explicitly modeled. In particular, canonization is sensitive to missed detections or spurious detections in the pontifical feature. For instance, in the illustrative case just discussed, in the presence of noise one would have different realization produce different numbers and location of local minima. This can be minimized by defining a scale-space of features, rather than considering the signal only at the resolution defined by the sample frequency of the sensor [10], and [19] for a more thorough discussion on this issue.

6 Discussion

The impact of the system-theoretic approach to dynamic data analysis has yet to be felt in important areas of applications such as data mining or computer vision. In order for this to happen, more general and flexible models have to be introduced. In this work we have made a modest step in this direction, by introducing Time Warping under Dynamic Constraints (TWDC), a method to compare time series that respects their dynamics. We have also introduced, albeit in a purely formal manner, a correlation kernel between processes that would respect their dynamic structure, if one could afford the time to compute the expectation with respect to the joint density of the driving noises. Finally, we have illustrated how one could, in principle, construct time-deformation-invariant statistics from time series to arrive at canonical representatives that are not affected by nuisances. These hold the promise for more efficient comparison, for each sequence can be pre-processed and comparison is performed by the simple computation of an $\mathbb{L}^2$ norm.

Acknowledgments

This work has benefited from extended discussions with Alessandro Chiuso, Andrea Vedaldi, Gregorio Guidi, Michalis Raptis, and René Vidal. Anything good there might be in it, it has been influenced – directly or indirectly – by Giorgio Picci's work. All the rest is sole responsibility of the author. The support of AFOSR and ONR is gratefully acknowledged.

References

1. P. Billingsley. *Convergence of Probability Measures*. Wiley, 1968.
2. A. Bissacco, A. Chiuso, and S. Soatto. Classification and recognition of dynamical models: the role of phase, independent components, kernels and optimal transport. *IEEE Trans. Pattern Anal. Mach. Intell.*, in press, 2007.
3. B. Giannakis and J. Mendel. Identification of nonminimum phase systems using higher order statistics. *IEEE Trans. on Acoustic, Speech and Signal Processing*, 37(3):360–377, 1989.
4. U. Grenander. *General Pattern Theory*. Oxford University Press, 1993.
5. A. Isidori. *Nonlinear Control Systems*. Springer Verlag, 1989.
6. G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211, 1973.
7. H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2004.
8. E. J. Keogh and M. J. Pazzani. Dynamic time warping with higher order features. In *Proceedings of the 2001 SIAM Intl. Conf. on Data Mining*, 2001.
9. A. Kirsch. An introduction to the mathematical theory of inverse problems. *Springer-Verlag, New York*, 1996.
10. T. Lindeberg. Scale space for discrete signals. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(3):234–254, 1990.
11. A. Lindquist and G. Picci. The stochastic realization problem. *SIAM J. Control Optim. 17*, pages 365–389, 1979.
12. C. F. Martin, S. Sun, and M. Egerstedt. Optimal control, statistics and path planning. 1999.
13. R. Martin. A metric for arma processes. *IEEE Trans. on Signal Processing*, 48(4):1164–1170, 2000.
14. P. J. Olver. *Equivalence, Invariants and Symmetry*. Cambridge University Press, 1995.
15. J. O. Ramsey and B. W. Silverman. *Functional Data Analysis*. Springer Verlag, 2005.

16. S. Soatto and A. Yezzi. Deformotion: deforming motion, shape average and the joint segmentation and registration of images. In *Proc. of the Eur. Conf. on Computer Vision (ECCV)*, volume 3, pages 32–47, 2002.
17. S. Soatto, A. J. Yezzi, and H. Jin. Tales of shape and radiance in multiview stereo. In *Intl. Conf. on Comp. Vision*, pages 974–981, October 2003.
18. A. Vedaldi and S. Soatto. Features for recognition: viewpoint invariance for non-planar scenes. In *Proc. of the Intl. Conf. of Comp. Vision*, October 2005.
19. A. Vedaldi and S. Soatto. Viewpoint induced deformation statitics and the design of viewpoint invariant features: singularities and occlusions. In *Eur. Conf. on Comp. Vision (ECCV)*, pages II–360–373, 2006.
20. A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury. The function space of an activity. In IEEE, editor, *Proc. IEEE Conf. on Comp. Vision and Pattern Recogn.*, 2006.
21. S.V.N. Vishwanathan, R. Vidal, and A. J. Smola. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *International Journal of Computer Vision*, 2005.
22. A. Yezzi and S. Soatto. Stereoscopic segmentation. In *Proc. of the Intl. Conf. on Computer Vision*, pages 59–66, 2001.