Classification and Recognition of Dynamical Models:

The Role of Phase, Independent Components, Kernels and Optimal Transport

UCLA CSD-TR 060020

Alessandro Bissacco Alessandro Chiuso Stefano Soatto

Abstract – **We address the problem of performing decision tasks, and in particular classification and recognition, in the space of dynamical models in order to compare time series of data. Motivated by the application of recognition of human motion in image sequences, we consider a class of models that include linear dynamics, both stable and marginally stable (periodic), both minimum and non-minimum phase, driven by non-Gaussian processes. This requires extending existing learning and system identification algorithms to handle periodic modes and non-minimum phase behaviour, while taking into account higher-order statistics of the data. Once a model is identified, we define a kernel-based cord distance between models that includes their dynamics, their initial conditions as well as input distribution. This is made possible by a novel kernel defined between two arbitrary (non-Gaussian) distributions, which is computed by efficiently solving an optimal transport problem. We validate our choice of models, inference algorithm, and distance on the tasks of human motion synthesis (sample paths of the learned models), and recognition (nearest-neighbor classification in the computed distance). However, our work can be applied more broadly where one needs to compare historical data while taking into account periodic trends, non-minimum phase behaviour, and non-Gaussian input distributions.**

Index Terms –**System Identification, Blind Deconvolution, Non-minimum Phase, Distance, Kernel, Hammerstein, Transport, Wasserstein, Non-Gaussian, Time Series.**

1 Introduction

Our goal is to perform decision tasks, including detection and recognition, in the space of dynamical models. For example, if we view a walking person as a dynamical system, we are interested in detecting her in an image sequence, recognizing her gait and possibly her identity. Endowing the space of dynamical models with a metric and a probabilistic structure is a long-standing problem because, even for linear models, such a space is highly non-linear. In comparing dynamical models one has to consider all of their components: The states and their *dynamics*, the *measurement maps*, the *initial conditions*, and the *inputs* or *noise* distributions. Different components may play different roles depending on the application; for instance, one may want to discard the transient behavior or the input distribution, but it is important to have machinery to account for all if needed.[1] We will concentrate on a class of models that is sufficiently general to capture many of the applications of interest in dynamic vision, and at the same time tractable in the sense of yielding, for the most part, closed-form (i.e. non-iterative) inference algorithms. As we explain in the next paragraph these are *marginally stable, non-minimum phase linear models with non-Gaussian inputs.*

Linear non-Gaussian models (Hammerstein)

Let $y(t) \in \mathbb{R}^m,\ t = t_0, t_1, \dots$ be the measured signal, sampled at discrete time intervals. Our goal is to describe its temporal behavior via a dynamical model. Under mild assumptions [23] $y(t)$ can be expressed as an instantaneous function of some "state" vector $x(t) \in \mathbb{R}^n$ that evolves in time according to an ordinary differential equation (ODE) driven by some deterministic or stochastic "input." In general, both the *measurement map* from $x(t)$ to $y(t)$ and the *state equation* that describes the ODE are non-linear, and complex dynamic phenomena such as hysteresis, phase transitions, turbulence, or delays require dedicated analytical tools. However, many non-linearities can be eliminated by proper choice of coordinates or immersion

[1] For the case of human gaits, one can think of the periodic dynamics as limit cycles generating nominal input trajectories, the stable dynamics governing muscle masses and activations, the initial conditions characterizing the spatial distribution of joints, and the input depending on the actual gait, the terrain, and the neuromuscular characteristics of the individual.

into a higher-dimensional spaces [26]. Indeed one can test the hypothesis that a given time series of measurements come from a linear model [21], as we do in Sect. 5 for the case of human motion. We will therefore restrict our attention to linear dynamical models of the type

$$\begin{cases} x(t+1) = Ax(t) + v(t) \quad x(t_0) = x_0 \\ y(t) = Cx(t) + w(t) \quad \{v(t), w(t)\} \overset{IID}{\sim} q(\cdot) \end{cases} \tag{1}$$

where $v(t)$ and $w(t)$ are stochastic inputs jointly described by the density $q(\cdot)$.[2] They can be though of as errors that compound the effects of unmodeled dynamics, linearization residuals, calibration errors and sensor noise. For this reason they are often collectively called input (state) and output (measurement) "noises." The density $q(\cdot)$ can be *Gaussian* or *non-Gaussian*. Given the linearity assumption on the model, the Gaussianity of the noise can be easily tested. For the case of human gaits, this reveals strongly non-Gaussian statistics (see Fig. 1). For reasons that will become clear shortly, we assume that the noise process is temporally independent, or *white*. Since one can interpret a white non-Gaussian independent and identically distributed (IID) process as a Gaussian one filtered through a static non-linearity, we are left with considering so-called *Hammerstein models*, that are linear models with static input non-linearities [18].

In the system identification literature [29, 41] it is customary to consider (1) as a description of the second-order statistics of the data (mean and covariance sequences). Indeed, there is an entire equivalence class of models of the form (1) that yield the same mean and covariance [27] and, therefore, one usually chooses the model that is *stable* and *minimum-phase*, i.e. the one with both poles and zeros inside the complex unit circle [29, 41, 34]. One can easily allow *marginally stable* modes (i.e. poles *on* the unit circle) provided that they are not "disturbed" by the input noise $v(t)$.[3] These describe periodic modes of the signal that are useful in many

[2]Deterministic inputs can be easily accounted for as a limiting case, and will therefore not be considered here.

[3]More precisely, the simple eigenvalues of A on the unit circle must correspond to unreachable components of the state, see later sections for more details.

applications, for instance in the analysis of gaits, as we describe in the next section.

However, when the inputs are non-Gaussian, models that are equivalent up to second-order may not, in general, produce the same higher-order statistics. In particular all (marginally) stable models with matching second-order statistics differ in their *phase,* which depends on the location of their zeros. Therefore, in our models we will have to forgo the minimum-phase assumption. This is appropriate for the case of human motion, where the underlying system is only marginally stable (there is a strong periodic component) and non-minimum phase: In fact, the body is a collection of inverted penduli, and the inverted pendulum is the prototypical example of non-minimum phase system.

Thus, the models we will consider are *marginally stable, non-minimum phase* of the form (1) with *non-Gaussian input and output noises.*

Goals and contributions

For this class of models we are interested in performing *blind identification*, i.e. to infer the model $M \doteq \{A, C, x_0, q(\cdot)\}$ from a time series $\{y(t_1), \dots, y(t_n), \dots\}$, and then to perform *classification* by endowing the space of such models with a distance $d(M_1, M_2)$. We wish to solve these tasks as much as possible without resorting to computationally intensive iterative optimization. Unfortunately, the current literature does not provide a solution to these tasks: Most algorithms either assume minimum-phase stable models [34] or involve iterative optimization [12, 42]. Similarly, model comparison is currently performed using spectral information [13, 31] that does not consider phase information, inputs or initial conditions. Recent work using kernels to compare dynamical models [44] only considers the inputs if they are identical (in which case the distance depends only on the transient behavior) or if they are independent (in which case the distance is not affected by the input). So, in order to accomplish our task of performing decisions in the space of models we will have to make several contributions to the state of the art:

- Develop closed-form system identification algorithms (ID) for linear models with *periodic modes*, extending the work on subspace algorithms [34],

- develop closed-form ID for *non-minimum phase* models, extending [34],
- develop closed-form *blind ID* of Hammerstein models with *non-Gaussian* inputs, extending the work of [18],
- introduce a novel *distance* between models that allows proper comparison of the *input distributions*. We will do so using *kernels*, thus extending the work of [44].

We will test our algorithms on the problems of human gait synthesis (ID) and recognition (distance). Along the way, we will

- point out relationships between our approach and traditional *subspace ID*, *independent component analysis* (ICA), *kernels*, and the problem of *optimal transport* and the associated *Wasserstein distance*.

We feel that each item in isolation is a useful contribution to the field, and in particular the introduction of transport-based kernels in Sect. 4.2 that could find application in other domains outside gait analysis. More importantly, we feel that the ensemble yields a unified picture and proposes a coherent method that allows classification of linear systems without many of the restrictions implicitly or explicitly present in much of the existing work.

Main ideas

The space of dynamical models, even linear ones, is strongly non-linear, and few attempts have been made to compute geodesic distances that would take the geometry of the space into account. More common is to define cord-distances that do not come from a metric. The most recent example is the work of Smola and coworkers [44] that define a kernel, i.e. an inner product in the embedding space of the output time series, and use that to define a distance. The kernel can be decomposed into a sum of terms, allowing the user to discount undesired elements (e.g. input, initial condition etc.) from the distance. Unfortunately, in order to compare two models M_1, M_2, the method proposed in [44] requires knowledge of the *joint density* of the noises, i.e. $p(v_1, w_1, v_2, w_2)$, which is seldom available.

The main idea of our method is to *identify a model that generates the same output statistics (of all orders) of the original system, but that has a canonical input* that is strongly white and with independent components. Then all the information content of the input is transferred to the model, that becomes non-linear (Hammerstein). One can then proceed to define a kernel in a manner similar to [44], but extended to take into account the non-linearity. This can be done by solving an *optimal transport* problem which, given a finite amount of data, can be done in closed-form.

Identification of the model can be conceptually broken down into steps: First, without loss of generality, we transform (1) in the form:

$$\begin{cases} x(t+1) = Ax(t) + Kn(t) \\ y(t) = Cx(t) + n(t). \end{cases} \tag{2}$$

Under our assumptions the noise $n(t)$ is temporally (strongly) white, and its components are *weakly* independent (uncorrelated). Note that, in general, the system above is non-minimum phase. This step shall be described in Sections 2 and 3. Then we *normalize* this model to make the components of the noise *strongly independent*. This is equivalent of performing *independent component analysis* (ICA) $n(t) = D\epsilon(t)$, yielding a model of the form

$$\begin{cases} x(t+1) = Ax(t) + B\epsilon(t) \\ y(t) = Cx(t) + D\epsilon(t) \end{cases} \tag{3}$$

with $B = KD$ and the components of ϵ are independent zero-mean unit-variance IID processes

$$\epsilon(t) = \begin{bmatrix} \epsilon_1(t) & \epsilon_2(t) & \cdots & \epsilon_m(t) \end{bmatrix}^\top \quad , \quad \epsilon_i(t) \overset{IID}{\sim} q_i(\epsilon_i) \quad , \quad E[\epsilon(t)\epsilon(t)^\top] = I. \tag{4}$$

and ϵ can be written in terms of a canonical (e.g. uniform, or Gaussian) noise u:

$$\begin{cases} x(t+1) = Ax(t) + Bf(u(t)) \quad x(t_0) = x_0 \\ y(t) = Cx(t) + Df(u(t)) \end{cases} \tag{5}$$

Now we can define a kernel, and therefore a distance, on the representation of the model

$M = \{A, B, C, D, x_0, f\}$ that takes into account the dynamics, measurement map, initial conditions, *and* the input statistics of the model. The hypothetical experiment to compare two models consists of randomly generating a scalar IID sequence distributed uniformly in [0 1], feeding it to the two models, and then compare their outputs.

2 Linear Models for Non-Gaussian Stationary Processes

In this section we introduce a linear dynamical system representation of stationary non-Gaussian processes with periodic modes. We use the notation $\hat{y}(t|t-1)$ to denote the best (minimum variance) linear predictor of $y(t)$ given its past history $\{y(t-1), y(t-2), ...\}$ [29]. It is well known (first part of Wold's decomposition theorem [38, 35]) that every stationary random process $y(t)$ can be decomposed into two parts

$$y(t) = y_d(t) + y_s(t) \tag{6}$$

where $y_d(t)$ is a *purely deterministic* (PD) process which can be predicted exactly as a linear combination of its past (i.e. $y_d(t) = \hat{y}_d(t|t-1)$) , and $y_s(t)$ is a *purely non-deterministic* (PND) process (or "purely stochastic," hence the choice of subscript s), uncorrelated from $y_d(t)$, for which the one step ahead prediction error $y_s(t) - \hat{y}_s(t|t-1)$ is different from zero in mean square. From Wold's decomposition theorem [38], the PND part[4] can be given an infinite moving average representation of the form $y_s(t) = \sum_{\tau=0}^{\infty} H(\tau)e(t-\tau)$, where $H(\tau)$ is a sequence of matrices such that $H(0) = I$, $\sum_{\tau=0}^{\infty} |H_{ij}(\tau)|^2 < \infty$ and $e(t)$ is the *innovation process* $e(t) \doteq y_s(t) - \hat{y}_s(t|t-1)$; $e(t)$ is uncorrelated $E[e(t)e^{\top}(s)] = 0$ for $t \neq s$.

Our tests in Sect. 5 suggest that for human gait data the linearity assumption leads to a good approximation of higher order statistics as well. This induces us to postulate a decomposition of the same form:

$$y_s(t) = \sum_{\tau=0}^{\infty} G(\tau)n(t-\tau) \tag{7}$$

[4]Usually only zero-mean processes are considered. However, the mean can be thought of as a PD component and hence included in $y_d(t)$.

where $G(0) = I$, $\sum_{\tau=0}^{\infty} |G_{ij}(\tau)|^2 < \infty$ and $n(t)$ is a strongly white (independent) process. Note that in general $H(\tau)$ needs not be equal to $G(\tau)$, as we shall discuss later. It is possible to show that the PD component $y_d(t)$ can be represented as the superposition of (possibly infinitely many) sinusoidal signals. However, from a practical standpoint, we can assume that $y_d(t)$ is the superposition of a finite number of sinusoids and hence can be represented, for $t > t_0$, as the output of an autonomous system of state dimension n_d

$$\begin{cases} x_d(t+1) &= A_d x_d(t) \\ y_d(t) &= C_d x_d(t) \end{cases} \tag{8}$$

with the constraint that A_d has simple eigenvalues on the unit circle. Without loss of generality the pair (A_d, C_d) can be taken to be observable [8]. From stationarity of y_d, $x_d(t)$ is also stationary and $P_d = \mathrm{Var}\{x_d(t)\}$ satisfies the homogeneous Lyapunov equation $P_d = A_d P_d A_d^\top$. Since the choice of basis in the state space is arbitrary, one can choose it so that $P_d = I$; with this canonical choice, we have

$$A_d A_d^\top = I \tag{9}$$

showing that, in this particular basis, A_d needs to be orthogonal.

Similarly, it is possible to give a state space realization to the representation (7) in the form

$$\begin{cases} x_s(t+1) &= A_s x_s(t) + K_s n(t) \\ y_s(t) &= C_s x_s(t) + n(t) \end{cases} \tag{10}$$

where $x_s(t) \in \mathbb{R}^{n_s}$. Defining the aggregate state $x(t) = [x_d^\top(t)\ x_s^\top(t)]^\top$, $x(t) \in \mathbb{R}^n$, we obtain

a generative model of the the stationary process $y(t) \in \mathbb{R}^m$ in state-space form (2) with

$$
\begin{aligned}
A &= \begin{bmatrix} A_d & 0 \\ 0 & A_s \end{bmatrix} \qquad |\lambda_i(A_d)| = 1,\ i = 1,..,n_d \ \ |\lambda_j(A_s)| < 1, \ \ j = 1,..,n_s \\
K &= \begin{bmatrix} 0 \\ K_s \end{bmatrix} \quad C = \begin{bmatrix} C_d & C_s \end{bmatrix} \quad x(t) = \begin{bmatrix} x_d(t) \\ x_s(t) \end{bmatrix}, \ x_d(t) \in \mathbb{R}^{n_d}, \ x_s(t) \in \mathbb{R}^{n_s} \\
&E\left[n(t)\right] = 0 \quad , \quad E\left[n(t)n(t)^\top\right] = R \quad , \qquad n(t) \text{ IID.}
\end{aligned} \tag{11}
$$

where $x_d(t)$ and $x_s(t)$ are the deterministic and stochastic components of the state corresponding to (6) and $x_0 = \begin{bmatrix} x_{0d}^\top & x_{0s}^\top \end{bmatrix}^\top$ is the initial condition. Notice that the only assumption we have on the input process $n(t)$ is that its samples are identically distributed and statistically independent.

Standard ID techniques provide the minimum-phase system estimate, that is the one that has zeros inside the unit circle $|\lambda(A - KC)| \leq 1$. However, the minimum-phase assumption typically does not hold for articulated mechanical systems such as the human body [33]. We do not assume that the underlying model is minimum-phase and, exploiting the fact that data are *not* Gaussian, we use higher-order *temporal* statistics to estimate a linear system of the form (11) (minimum or non-minimum phase) that best matches the observations.

Finally, we extend the model (11) to include higher-order *spatial* statistics by assuming the instant mixing model (4) for the input process $n(t) = D\epsilon(t)$. That is, $n(t)$ is a linear transformation of the process $\epsilon(t)$ whose components have non-Gaussian distributions q_i and are both temporally and spatially statistically independent. Estimating the mixing matrix D and the IID processes $\epsilon_i(t)$ from $n(t)$ is a standard independent component analysis (ICA) [14], for which efficient learning algorithms exist [6].

Our goal is that of finding optimal, in some sense, estimates of the parameters of the combined model $(2, 11, 4)$ that is the matrices A, K, C, the initial state x_0, the mixing matrix D and the input distributions q_i.

3 Inference criteria and learning algorithms

In order to estimate the parameters of the dynamical model (11) and the input distribution (4) we propose a two-stage learning approach. Our method differs from common blind deconvolution/system identification approaches in that *it handles critically stable systems, i.e. system with poles on the unit circle.*[5] Another main distinction from common gradient-based approaches such as [12, 42] is that we do not require solving computationally expensive high-dimensional optimization problems with many local minima. We propose a non-iterative suboptimal approach that provides a direct estimate in closed form. This approximate solution can also be used as an initial guess for any gradient-based learning algorithm if so desired.

First we construct the set of linear systems $M_l = \{A, K_l, C, R_l\}$, $l = 1, \cdots, L$ which match the second order statistics of the data. This is known as the *stochastic realization* problem [27]. All models matching the second order statistics share the same A, C matrices, that can be estimated by subspace identification techniques [34]; now we shall extend subspace techniques to handle the presence of PD components $y_d(t)$.

The set of matrices K_l, R_l, $l = 1, \cdots, L$, can be obtained solving a Riccati equation once (A, C) have been computed [28]. Then for each system we compute its inverse and use it to estimate the input process $n_l(t)$ and the initial state x_0 from a realization of $y(t)$. An independence test based on higher-order statistics of $n_l(t)$ is used to select the system M_l which best matches the observed process $y(t)$. Finally the mixing matrix and the input distributions in (4) are estimated using an efficient independent components analysis algorithm [6].

3.1 Estimating Second-Order Statistics

Standard approaches to estimation of periodic signals corrupted by white noise, such as MUSIC [39], ESPRIT [37] and related algorithms [24, 17], discard non-periodic modes. However,

[5] Our model (11) should not be confused with cointegrated models used in econometrics (see e.g. [3]) where the state components corresponding to eigenvalues on the unit circle are reachable, i.e. are affected by the noise $n(t)$, implying non-stationarity of the output process $y(t)$.

such modes are of paramount importance in our application and, therefore, these algorithms cannot be applied directly. We claim that the standard subspace procedure can be modified so as to infer models of the form (2, 11); the methodology we propose shall also provide a natural criterion to estimate the number of PD components. The procedure is composed of two steps: First, standard subspace identification [43, 34] is applied to the signal $y(t)$ as if there was no PD component. It is possible to show, but beyond the scope of this paper, that as the number of data grows, the algorithm guarantees a consistent estimate of the parameters of the minimum-phase system describing the data. However, with finite data we get an estimate of the system matrices which, with an appropriate choice of basis T, are of the form:

$$T^{-1}\hat{A}T = \begin{bmatrix} \hat{A}_d & 0 \\ 0 & \hat{A}_s \end{bmatrix} \quad T^{-1}\hat{K} = \begin{bmatrix} \hat{K}_d \\ \hat{K}_s \end{bmatrix} \quad \hat{C}T = \begin{bmatrix} \hat{C}_d & \hat{C}_s \end{bmatrix} \tag{12}$$

where in general neither $|\lambda(\hat{A}_d)| = 1$ nor $\hat{K}_d = 0$. Therefore, we adopt a second step to guarantee that both $|\lambda(\hat{A}_d)| = 1$ and $\hat{K}_d = 0$. In order to do so, we need to review the basic steps performed in subspace identification which we shall modify to our purpose.

Let us define $Y_t \doteq \frac{1}{\sqrt{N}}[y(t), y(t+1), .., y(t+N-1)]$ and $X_t \doteq \frac{1}{\sqrt{N}}[x(t), x(t+1), ...x(t+N-1)]$. The number of columns N here depends on the number of data available. As discussed in [11], subspace identification can be seen as a two step procedure as follows:

1. Construct a basis $\hat{X}_t$ for the state space via suitable projection operations on data sequences (Hankel data matrices).

2. Given (coherent) bases for the state space at time t, $(\hat{X}_t)$, and $t+1$, $(\hat{X}_{t+1})$, compute the least-squares solution to

$$\begin{cases} \hat{X}_{t+1} \simeq A\hat{X}_t + K(Y_t - \hat{C}X_t) \\ Y_t \simeq C\hat{X}_t \end{cases} \tag{13}$$

Since we do not need to modify the first step, we refer the reader to [43, 11] for details. As for the second step, we need to solve it while enforcing the constraint that the estimated A

matrix has n_d eigenvalues[6] lying on the unit circle. In order to do so, we follow the steps:

1. solve (13) in the least-squares sense obtaining $(\hat{A}, \hat{C}, \hat{K})$;

2. compute the eigenvalue decomposition of $\hat{A}$ and let T be the change of basis; estimate n_d as the number of eigenvalues of $\hat{A}$ which are "close" to the unit circle and whose corresponding eigenspace is "almost unreachable"[7] using the input matrix $\hat{K}$. Without loss of generality we assume the first n_d elements of the state span this "almost" unreachable subspace so that, with this choice of basis (12) holds with $\hat{K}_d \simeq 0$ and $|\lambda(\hat{A}_d)| \simeq 1$. It is possible to devise theoretically sound statistical tests for performing this decision using recent results on the asymptotic properties of subspace estimators [2, 9] that go beyond our scope here. We only note that this corresponds to estimating the subspace of the state space which generates the PD components, including its dimension n_d. In fact, with this choice of basis, the state matrix $\hat{Z}_t \doteq T^{-1}\hat{X}_t$ can be partitioned as follows:

$$\hat{Z}_t = T^{-1}\hat{X}_t = \begin{bmatrix} \hat{Z}_t^d \\ \hat{Z}_t^s \end{bmatrix} \quad \hat{Z}_{t+1} = T^{-1}\hat{X}_{t+1} = \begin{bmatrix} \hat{Z}_{t+1}^d \\ \hat{Z}_{t+1}^s \end{bmatrix}.$$

   Without loss of generality it is possible to chose T so that $\hat{Z}_t^d(\hat{Z}_t^d)^\top = I$. Note also that this implies $\hat{Z}_{t+1}^d(\hat{Z}_{t+1}^d)^\top \simeq I$. This shall be useful later on.

   It is straightforward to show that solving the least-squares problem (13) with this new choice of basis $\hat{Z}_t$ corresponds *exactly* to changing basis in the estimated state matrices $\hat{A}$, $\hat{K}$, $\hat{C}$, i.e. $\hat{A} = T\hat{A}_T\hat{T}^{-1}$, $\hat{K} = T\hat{K}_T$, $\hat{C} = \hat{C}_T T^{-1}$ where

$$\begin{aligned} \hat{C}_T &\doteq \arg\min_{C_T} \|Y_t - C_T\hat{Z}_t\|_F \\ \hat{A}_T, \hat{K}_T &= \arg\min_{A_T, K_T} \|\hat{Z}_{t+1} - A_T\hat{Z}_t - K_T(Y_t - \hat{C}_T\hat{Z}_t)\|_F \end{aligned} \tag{14}$$

   Note also that, using the fact that the rows of $Y_t - \hat{C}_T\hat{Z}_t$ are orthogonal to the rows of $\hat{Z}_t$, the problem of estimating $\hat{A}_T$ can be further simplified to:

[6] Recall that n_d is not known *a priori* and should be estimated from data. Like all model selection techniques, there is a design parameter involved in our procedure.

[7] This means that the restriction of the pair $(\hat{A}, \hat{K})$ to the desired subspace is nearly unreachale.

$$\hat{A}_T \;=\; \arg\min_{A_T} \|\hat{Z}_{t+1} - A_T \hat{Z}_t\|_F \tag{15}$$

Since, by construction (see equations (12), (14)), $\hat{A}_T$ has a block diagonal structure $\hat{A}_T = \mathrm{diag}\{\hat{A}_d, \hat{A}_s\}$, the matrices $\hat{A}_d$, $\hat{A}_s$ can be obtained via:

$$\hat{A}_d = \arg\min_{A_d} \|\hat{Z}^d_{t+1} - A_d \hat{Z}^d_t\|_F \qquad \hat{A}_s = \arg\min_{A_s} \|\hat{Z}^s_{t+1} - A_s \hat{Z}^s_t\|_F \tag{16}$$

Therefore, with this choice of basis, we can decouple the identification of the PD and PND components, making it easier to introduce the constraints $|\lambda(A_d)| = 1$ and $K_d = 0$.

3. Solve the constrained least-squares problem: $\arg\min_{|\lambda(A_d)|=1} \|\hat{Z}^d_{t+1} - A_d \hat{Z}^d_t\|_F$. Note that, in the ideal case, $\hat{Z}^d_{t+1} = A_d \hat{Z}^d_t$ so that $I = \hat{Z}^d_{t+1}(\hat{Z}^d_{t+1})^\top = A_d \hat{Z}_t (\hat{Z}^d_t)^\top A_d^\top = A_d A_d^\top$ where the first and last equality follows from the choice of basis which guarantees $\hat{Z}^d_t (\hat{Z}^d_t)^\top = I$. This implies that A_d needs to be an orthogonal matrix. This observation is the "sample version" of (9) for the second-order moments of $x_d(t)$. Therefore, we obtain the following matrix Procrustes problem

$$\hat{A}^c_d = \underset{A_d \in O(n_d)}{\arg\min} \|\hat{Z}^d_{t+1} - A_d \hat{Z}^d_t\|_F \tag{17}$$

that can be easily solved using the singular value decomposition of $\hat{Z}^d_{t+1}(\hat{Z}^d_t)^\top$ [19, 22]:

$$\hat{A}^c_d = U_a V_a^\top \quad , \quad \hat{Z}^d_{t+1}(\hat{Z}^d_t)^\top = U_a \Sigma_a V_a^\top \tag{18}$$

4. The remaining system parameters are computed as follows:

   (a) Using the estimated $\hat{A}^c_d$ of the previous step and $\hat{C}_d$ from the block decomposition (12) define an estimate of the "deterministic" observability matrix $\hat{\Gamma}^\top_d(N) \doteq \left[\hat{C}^\top_d \; (\hat{A}^c_d)^\top \hat{C}^\top_d \ldots \left((\hat{A}^c_d)^{N-1}\right)^\top \hat{C}^\top_d\right]$; define also $Y \doteq [y^\top(1), .., y^\top(N)]^\top$. Estimate (in the least squares sense) the initial condition $\hat{x}_0$ from $Y \simeq \hat{\Gamma}_d x_0 + Y_s$, minimizing the norm of Y_s. This allows us to remove the PD component and extract

$$\hat{Y}_s \doteq Y - \hat{\Gamma}_d(N)\hat{x}_0 = (I - \hat{\Gamma}_d(N)(\hat{\Gamma}^\top_d(N)\hat{\Gamma}_d(N))^{-1}\hat{\Gamma}_d(N))Y.$$

From the vector $\hat{Y}_s \doteq [\hat{y}_s^\top(1), .., \hat{y}_s^\top(N)]^\top$ estimate the "stochastic" parameters $\hat{A}_s, \hat{C}_s$, $\hat{\Lambda}_s \simeq \text{Var}\{y_s(t)\}$ and $\hat{G}_s \simeq E\left[y_s(t)x_s^\top(t+1)\right]$ using [43]. This pre-filtering step, used to remove the PD component, is similar to the prefiltering step used in the orthogonal decomposition algorithm described in [36, 10].

(b) Estimate $\hat{R}_l$ and $(\hat{K}_s)_l$ solving the Riccati equation as described in the technical report [5] [8]. The index l refers to the different solutions of the Riccati equation.

The estimated (constrained) state matrices are then given by

$$\hat{A}^c = \begin{bmatrix} \hat{A}_d^c & 0 \\ 0 & \hat{A}_s \end{bmatrix} \quad \hat{K}_l^c = \begin{bmatrix} 0 \\ (\hat{K}_s)_l \end{bmatrix} \quad \hat{C}^c = \begin{bmatrix} \hat{C}_d & \hat{C}_s \end{bmatrix}$$

and noise variance $E[n_l(t)n_l^\top(t)] \simeq \hat{R}_l$.

Without delving into details (see [5] for derivations), the discrete Riccati equation provides a finite number L of solutions corresponding to picking, for each zero-pair of the system, either the zero inside the unit circle or its conjugate reciprocal. This allows us to efficiently recover all the systems (11) generating the same second-order statistics of $y(t)$. Each solution $l = 1, \cdots, L$ corresponds to a different factorization of the power spectrum $S_y(z)$, and is given by the same internal dynamics A, C but different input-related matrices $\{K_l, R_l\}$. In the end, we have a set of candidate models $M = \{A, K_l, C, R_l\}$ (11) which are parametric representations of the second-order statistics of the data. In order to choose among these the one which most closely matches the statistics of $y(t)$, we need to investigate higher-order temporal dependencies.

3.2 Temporal Independence and Phase Estimation

In this section we deal with the problem of estimating the phase of (11), that is selecting from a finite set of linear systems $M_l = \{A,\ K_l,\ C, R_l\}$ $l = 1, \ldots, L$ the one which best matches the temporal statistics of the process $y(t)$. Our approach is similar to the one proposed in [7] for

[8] The input-to-state matrix K is computed via the Riccati equation. Only the PND components enter in this calculation while we force $\hat{K}_d = 0$ in the estimated model.

scalar signals, in that we use the inverse system $M_l^{-1} = \{A - K_l C,\ K_l,\ -C\}$ to estimate the white input $n_l(t)$. However, we cannot assume that the components of $n_l(t)$ are independent, therefore we cannot use a standard contrast function (e.g. kurtosis or negentropy)[9] to select the best match, as done in [7].

If we choose a model M_l with phase structure different from the underlying true system, we introduce higher-order temporal correlations in the estimated input $n_l(t)$, equivalently to filtering the true input process $n(t)$ with an all-pass filter. In this case $n_l(t)$ is still temporally uncorrelated, but its samples are no longer independent:

$$E[n_l(t)n_l(s)^\top] = 0 \quad , \quad p(n_l(t), n_l(s)) \neq p(n_l(t))p(n_l(s)).$$

Therefore, the best model for the dynamics of the process $y(t)$ is the one producing the input $n_l(t)$ most temporally independent. We introduce a simple measure of temporal independence for white processes using the third-order cumulant function[10], which for a zero-mean scalar process $x(t)$ is defined as:

$$\mathrm{cum}_{3x}(\tau_1, \tau_2) \doteq E\left[x(t)x(t-\tau_1)x(t-\tau_2)\right]. \tag{19}$$

It is easy to see that this function is symmetric, so it is sufficient to consider the nonredundant region $0 \leq \tau_2 \leq \tau_1$. If the samples $x(t)$ are independent, then the third-order cumulant is an impulse centered in zero: $\mathrm{cum}_{3x}(\tau_1, \tau_2) = \delta(\tau_1, \tau_2)E[x(t)^3] \quad , \quad x(t) \overset{IID}{\sim} p(x)$.
Given a realization of $x(t)$, we can measure its temporal independence using the normalized cross-correlation between the sample third cumulant and the impulse function:

$$\rho(x)^2 \doteq \frac{\hat{\mathrm{cum}}_{3x}^2(0,0)}{\sum_{i=-N}^{N}\sum_{j=-N}^{N}\hat{\mathrm{cum}}_{3x}^2(i,j)} \tag{20}$$

where ideally $N \to \infty$ but in practice it is sufficient to take the sum for a small number of

[9] A contrast function Ψ allows to discriminate a signal with independent components from its linear combinations by the property: $\Psi(Qn(t)) \leq \Psi(n(t))\ , \forall Q : QQ^T = I$. Unfortunately this cannot be used for selecting the most temporal independent among the candidate input processes $n_l(t)$.

[10] Notice that we tacitly assume that the input distribution is not symmetric, as it is the case for human motion applications. Otherwise, the third-order statistics would be identically zero, and it would be necessary to derive a temporal independence test from the fourth-order cumulants.

time lags. We can easily extend this independence score to a multivariate process $\mathbf{x}$ by taking the product of the scores computed independently on each component x_k:

$$\rho(\mathbf{x})^2 = \prod_{k=1}^{m} \rho(x_k)^2 = \prod_{k=1}^{m} \frac{\hat{\text{cum}}^2_{3x_k}(0,0)}{\sum_{i=-N}^{N} \sum_{j=-N}^{N} \hat{\text{cum}}^2_{3x_k}(i,j)} \tag{21}$$

Notice that deconvolving a non-minimum phase system in order to recover the input is a non-trivial process. The zeros outside the unit circle in the original system M become unstable poles of the inverse system M^{-1}. Then, unless we resort to approximations (such as [32]), the input signal cannot be computed in a causal fashion. In our application we can afford to operate offline, so we perform exact inversion. In brief, we decompose the inverse system in causal and acausal part by applying a similarity transformation that partitions A into stable and unstable components. Then we run the causal part forward in time and the acausal part backward in time. We repeat this process for every system M_l and we pick the one that produces the input $n_l(t)$ with the highest independency score (21).

With the inverse system, we also estimate the initial state x_0 associated to the realization $y(t)$. For nonminimum phase systems, in addition to the initial state x_0 we obtain a final state x_T which is the initial value for the a-causal part of the inverse system.

3.3 Estimating Spatial Statistics

Given the dynamical system M matching the dynamics of the gait process and the estimated IID input process $n(t)$, the last step is to use the higher order statistics to estimate the mixture model (4). This is a standard independent component analysis problem, and several approaches have been proposed for its solution. We applied the CubICA algorithm [6] which, unlike many others does not require manually tuning of parameters. It is based on joint diagonalization of third- and fourth-order cumulants.

The output of the ICA algorithm is the invertible mixing matrix D. We can recover the original signal $\epsilon(t)$ (up to sign and permutation ambiguities) by:

$$\epsilon(t) = D^{-1} n(t) \tag{22}$$

It is important to notice that the recovered input (22) matches the true input process (4) up to a sign ambiguity (we obtain the same $y(t)$ if we multiply each input $\epsilon_i(t)$ and the corresponding i-th column of D by -1) and an arbitrary permutations of its components. We will consider these ambiguities later when we define kernels for input processes.

Since $\epsilon(t)$ has independent components, we represent the distribution q_i of $\epsilon_i(t)$ by its sample histogram.

4 Kernels for Linear Systems

In this section we will define kernels for dynamical systems of the form $(3, 4)$ with input $\epsilon(t) \in \mathbb{R}^m$, state $x(t) \in \mathbb{R}^n$ and output $y(t) \in \mathbb{R}^m$. Here we only assume that the input is a unit variance IID stationary process with independent components. In the next section we will complete the model $(3, 4)$ to include the higher-order statistics of the process $y(t)$ by explicitly representing the distribution of the input components $\epsilon_i(t)$.

Given two models $\{A, B, C, D, x_0\}$, $\{A', B', C', D', x_0'\}$ and the unit-variance inputs $\epsilon(t), \epsilon'(t)$, we obtain the following outputs $y(t), y'(t)$:

$$\begin{aligned} y(t) &= CA^t x_0 + D\epsilon(t) + \sum_{i=0}^{t-1} CA^i B\epsilon(t-1-i) \\ y'(t) &= C'(A')^t x_0' + D'\epsilon'(t) + \sum_{i=0}^{t-1} C'(A')^i B'\epsilon'(t-1-i). \end{aligned} \tag{23}$$

If the inputs are Gaussian or they have the same higher-order statistics, we can define kernels between models (23) by assuming the same input:

$$\epsilon'(t) = \epsilon(t) \tag{24}$$

This allows us to compute the correlation matrix Σ between $y(t)$ and $y(t)'$ by marginalizing over the common noise $\epsilon(t)$:

$$\Sigma\left[(\{A, B, C, D, x_0\}, \{A', B', C', D', x_0'\}\right] \doteq E_\epsilon \left[\sum_{t=1}^{\infty} e^{-\lambda t} W y'(t) y(t)^\top\right] \tag{25}$$

where there is an exponential discounting term $e^{-\lambda t}$, $\lambda \geq 0$ and a user-defined symmetric weight matrix W. From (23) we have:

$$\begin{aligned} \Sigma\left[\{A,B,C,D,x_0\},\{A',B',C',D',x_0'\}\right] &= \Sigma\left[\{A,C,x_0\},\{A',C',x_0'\}\right] + \\ &+ \Sigma\left[D,D'\right] + \Sigma\left[\{A,B,C\},\{A',B',C'\}\right] \end{aligned} \tag{26}$$

where:

$$\Sigma\left[\{A,C,x_0\},\{A',C',x_0'\}\right] = \sum_{t=1}^{\infty} e^{-\lambda t} W C' (A')^t x_0' x_0^\top (A^\top)^t C^\top \tag{27}$$

$$\Sigma\left[D,D'\right] = E_\epsilon\left[\sum_{t=1}^{\infty} e^{-\lambda t} D' \epsilon'(t) \epsilon(t)^\top D^\top\right] \tag{28}$$

$$\Sigma\left[\{A,B,C\},\{A',B',C'\}\right] = E_\epsilon\left[\sum_{t=1}^{\infty} e^{-\lambda t} \sum_{i=0}^{t-1} C' (A')^i B' \epsilon(t-1-i)' \epsilon(t-1-i)^\top B^\top (A^\top)^i C^\top\right] \tag{29}$$

The correlation on the initial state (27) can be computed as in [44]:

$$\Sigma[\{A,C,x_0\},\{A',C',x_0'\}] = W C' V C^\top \quad , \quad V = e^{-\lambda} A' x_0' x_0^\top A^\top + e^{-\lambda} A' V A^\top \tag{30}$$

The correlation on the measurement noise (28) is:

$$\Sigma[D,D'] = (e^\lambda - 1)^{-1} W D' U D^\top \tag{31}$$

where $U \doteq E_\epsilon[\epsilon'(t)\epsilon(t)^\top]$, and since we assume the same (24) unit variance input (4) , we have $U = I$. Later we will use the input correlation matrix U to include the effect of the higher-order statistics of the input distributions. The correlation on the state noise (29) is:

$$\Sigma\left[\{A,B,C\},\{A',B',C'\}\right] = (e^\lambda - 1)^{-1} W C' \tilde{V} C^\top \quad , \quad \tilde{V} = B' U B^\top + e^{-\lambda} A' \tilde{V} A^\top \tag{32}$$

Then, from the output correlation matrix (25), we can define the trace kernel k_t as:

$$k_t(\{A,B,C,D,x_0\},\{A',B',C',D',x_0'\}) \doteq E_\epsilon\left[\sum_{t=1}^{\infty} e^{-\lambda t} y(t)^\top W y'(t)\right] = \tag{33}$$

$$= \mathrm{tr}\Sigma\left[\{A,B,C,D,x_0\},\{A',B',C',D',x_0'\}\right] \tag{34}$$

and the determinant kernel k_d as:

$$k_d(\{A,B,C,D,x_0\},\{A',B',C',D',x_0'\}) \doteq E_\epsilon \det\left[\sum_{t=1}^{\infty} e^{-\lambda t} y'(t) y(t)^\top\right] = \\ = \det\Sigma\left[\{A,B,C,D,x_0\},\{A',B',C',D',x_0'\}\right] \quad (35)$$

where, without loss of generality, we assume $\det W = 1$. Using the Binet-Cauchy theorem on compound matrices, in [44] it is shown that functions of the form $(34, 35)$ are dot products in an embedding space and they define positive definite kernels.

The trace kernels (34) provide some advantages with respect to the determinant kernels (35)[11]. There is also an interesting connection between trace kernels and the H_2 norm for linear systems commonly used in optimal control [12]. Given the autonomous system $M = \{A, C, x_0\}$, consider the Single Input Multiple Output linear system $\tilde{M} = \{e^{-\frac{\lambda}{2}}A,\ x_0,\ C,\ \mathbf{0}\}$. It is easy to see that the trace kernel $k_t(M, M)$ is the squared H_2 norm of $\tilde{M}$:

$$k_t(\{A,C,x_0\},\{A,C,x_0\}) = ||\{e^{-\frac{\lambda}{2}}A, x_0, C, \mathbf{0}\}||^2_{H_2}$$

Similar relations hold for the input related trace kernels.

The proposed kernels can be used to define a distance in the space of linear models. Let $M = \{A, B, C, D, x_0\}$, $M' = \{A', B', C', D', x_0'\}$ be two such models, then the kernel distance $d(M, M')$ is defined as:

$$d(M,M') = k(M,M) + k(M',M') - 2k(M,M') \quad (36)$$

This is a crucial ingredient to perform classification in the space of dynamical models.

[11] First they allow for more efficient computations in the case of high-dimensional data, since they can be computed from a $n \times n$ matrix derived from the dot product (33) instead of the determinant kernel which need to use the high-dimensional correlation matrix (27) (see [44] for details on calculations). When the measurements $y(t)$ are images, trace kernels are indeed the only computationally doable option. Another advantage of trace kernel compared to determinant kernels is that they do not introduce ambiguities on the sign of the correlation. For example if $y(t)$ has an even number of independent components and $y'(t) = -y(t)$, then the determinant kernel will give the same score as when the two processes are the same, while the trace kernel correctly identifies their negative correlation. Finally, the linearity of trace kernels allows to decompose the final result as the sum of the single contributions, that is initial state evolution (30) and input distribution $(31, 32)$.

[12] The H_2 norm of a stable system is defined as the Frobenius norm of its impulse response.

4.1 Initial State Alignment

Consider the case of two sequences generated by the same periodic process, observed with a phase delay τ. Using any identification algorithm we will estimate two systems $\{A, C, x_0\}$, $\{A', C', x_0'\}$ that, although representing the same signal, have very different initial states and consequently exhibit little similarity according to the kernels defined on the initial state correlation (27). Thus, in order to make the kernels invariant to delays, we introduce an alignment process that evolves the initial states x_0, x_0' for $\tau, \tau' \geq 0$ steps respectively so that the kernels $(34, 35)$ are maximized. That is, we define the aligned kernel k_a between initial states as:

$$k_a(\{A, B, C, D, x_0\}, \{A', B', C', D', x_0'\}) = \max_{(\tau,\tau')\in\mathcal{T}} k(\{A, B, C, D, A^\tau x_0\}, \{A', B', C', D', A'^{\tau'} x_0'\}) \tag{37}$$

where $\mathcal{T} \subset \mathbb{N}^2$ is the set of delays that we want our kernels to be invariant to. Unfortunately (37) is a system of exponential equations for which, to the best of our knowledge, no closed form solution is available.

Given that in many applications the period of dominant modes is short, we can afford to solve (37) by exhaustive search. Assuming that the aligning delay is no greater than T, we search for the maximum of (37) in the following set of $3T + 1$ delays:

$$\mathcal{T} = \{(0,0), (1,0), \cdots, (T,0), (0,1), \cdots, (0,T), (1,1), \ldots, (T,T)\} \tag{38}$$

When $\lambda = 0$, the symmetric delays $\tau = \tau' > 0$ can be omitted. They are however required in the general case in order for (37) to be a positive kernel.

4.2 Kernels for Arbitrary Input Distributions

In this section we will introduce the last crucial element of our approach, a kernel between arbitrary IID processes. Given a random variable x with density function p and cumulative distribution function $F : \mathbb{R} \mapsto [0, 1]$:

$$x \sim p(x) \quad , \quad F(a) = \int_{-\infty}^{a} p(x)dx = P[x \leq a] \tag{39}$$

we can use the quantile function F^{-1} (i.e. the inverse of the distribution function) to transform a uniform variate $u \in \mathbb{U}[0,1]$ into a random variable distributed according to F:

$$u \in \mathbb{U}[0,1] \;\rightarrow\; F^{-1}(u) \sim p(x). \tag{40}$$

Thus, we can define a kernel between pairs of (scalar) random variables x, x' having distributions F, F' as the correlation between the two random variables obtained by applying the same uniform u to the quantile functions F^{-1}, F'^{-1} :

$$k(x,x') = E_{u\sim\mathbb{U}[0,1]}[F^{-1}(u)F'^{-1}(u)] = \int_0^1 F^{-1}(u)F'^{-1}(u)du \tag{41}$$

Consider the linear manifold[13] $\mathcal{H}$ of random variables with zero mean and finite variance defined on the same probability space $(\Omega, \mathcal{F}, P)$. It is well known [38] that $\mathcal{H}$ can be made into an Hilbert space introducing the inner product $\langle x, x' \rangle \doteq E[xy]$. Then, (41) is a dot product and consequently a positive definite kernel. The distance induced by this kernel:

$$d_W(x,x') = k(x,x) + k(x',x') - 2k(x,x') = \int_0^1 |F^{-1}(u) - F'^{-1}(u)|^2 du \tag{42}$$

is known for probability distributions as Wasserstein, Mallows or Ornstein distance [30, 4]. It is more generally defined for two (possibly multidimensional) probability densities P and Q as $d_W(P,Q)^2 = \inf_J\{E_J[(X-Y)^\top(X-Y)] : \; (X,Y) \sim J, X \sim P, Y \sim Q\}$, where the infimum is taken over all the joint densities J which have marginals equal to P and Q. This distance represents the solution to the Monge-Kantorovich mass transfer problem, and can be interpreted as the minimum amount of work that is required to transport a mass of soil with distribution P to an excavation having distribution Q. For discrete distributions, the Wasserstein distance is equivalent to the Earth mover's distance, a metric commonly used for measuring texture and color similarities.

[13] I.e. the space of finite linear combinations of random variable in $(\Omega, \mathcal{F}, P)$, closed with respect to convergence in mean square.

From (42), we can compute the kernel between input distributions $k(x, x')$ from their Wasserstein distance $d_W(x, x')$. Using the change of variable $x = F^{-1}(u)$, it is easy to see that the kernel $k(x, x)$ gives the second moment of x: $k(x, x) = \int_0^1 |F^{-1}(u)|^2 du = \int_{-\infty}^{\infty} x^2 p(x) dx = E[x^2]$. Substituting (4.2) in (42) we obtain:

$$k(x, x') = \frac{1}{2}\left(E[x^2] + E[x'^2] - d_W(x, x')\right) \tag{43}$$

In case of zero-mean unit variance $E[x^2] = E[x'^2] = 1$, we have simply $k(x, x') = 1 - \frac{1}{2} d_W(x, x')$.

Although this expression is attractive, in the case of discrete distributions it is more efficient to compute the kernel by directly evaluating the integral (41). We can define a kernel between scalar IID processes $x(t), x'(t)$ as:

$$k(x(t), x'(t)) \doteq E_u\left[\sum_{t=1}^{\infty} e^{-\lambda t} F^{-1}(u(t)) F'^{-1}(u(t))\right] = (e^{\lambda} - 1)^{-1} E_u\left[F^{-1}(u) F'^{-1}(u)\right] \tag{44}$$

The kernel (44) can be extended to multivariate processes. Given an IID process $\epsilon(t) \in \mathbb{R}^m$ with independent components, it can be modeled as the output of its m quantile functions F_i^{-1} to m independent uniform processes $u_i(t)$, i.e. $\epsilon(t) = f(u(t))$, where $f(u(t)) = \begin{bmatrix} F_1^{-1}(u_1(t)) & \cdots & F_m^{-1}(u_m(t)) \end{bmatrix}$. Then, given two IID processes $\epsilon(t), \epsilon'(t) \in \mathbb{R}^m$ with independent components, they can be represented as outputs of two vector functions f, f' to the same input u:

$$\epsilon(t) = f(u(t)) \quad , \quad \epsilon'(t) = f'(\Pi(\sigma) u(t)) \tag{45}$$

where $\sigma \in S(m)$ (symmetric group of order m) is a permutation of the input representing correspondences between the elements of the two processes, i.e. each component i of ϵ is correlated with the component σ_i of ϵ' : $E[\epsilon_i \epsilon'_{\sigma_i}] \neq 0$, $E[\epsilon_i \epsilon'_j] = 0$ $j \neq \sigma_i$, and $\Pi(\sigma) = [\pi_{ij}]$ is the permutation matrix corresponding to σ, i.e. $\pi_{i\sigma_i} = 1$, $\pi_{ij} = 0$ $\forall j \neq \sigma_i$.

If the processes $\epsilon(t), \epsilon'(t)$ are inputs to a linear model of the form (2), the permutation σ represents the inherent ambiguity of the model, since we can obtain equivalent systems by rearranging the input elements $\epsilon_i(t)$ and the columns of the mixing matrix D. A direct consequence of this ambiguity is that the order in which the input components are recovered by the ICA algorithm is arbitrary. Additionally, there is a sign ambiguity, that is we can

change the sign of any $\epsilon_i(t)$ and of the corresponding i-th column of D (see Sect. 3.3).

Using (45), we can compute the correlation matrix U between vector processes $\epsilon(t), \epsilon'(t)$ with correspondences σ as:

$$\begin{aligned} U(\sigma) &\doteq E_u\left[\sum_{t=1}^{\infty} e^{-\lambda t} f'(\Pi(\sigma)u(t)) f(u(t))^\top\right] = \\ &= \Pi(\sigma)\begin{bmatrix} k(\epsilon_1(t), \epsilon'_{\sigma_1}(t)) & 0 & \cdots & 0 \\ 0 & k(\epsilon_2(t), \epsilon'_{\sigma_2}(t)) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & k(\epsilon_m(t), \epsilon'_{\sigma_m}(t)) \end{bmatrix} \end{aligned} \tag{46}$$

Given the correspondences σ, we can define the trace kernel between $\epsilon(t)$ and $\epsilon'(t)$ as:

$$k_t(\epsilon(t), \epsilon'(t); \sigma) = \mathrm{tr}\left(\Pi(\sigma)^\top |U(\sigma)|\right) = \sum_{i=1}^{m} |k\left(\epsilon_i(t), \epsilon'_{\sigma_i}(t)\right)| \tag{47}$$

where we use the absolute value of the correlation between input components to resolve the sign ambiguity. This is a symmetric positive function of the input distributions, therefore is a positive definite kernel [40]. If the correspondences σ are unknown, we can compute the optimal trace matching $\hat{\sigma}_t$ as the solution to the maximum-weight assignment problem defined by the $m \times m$ Gram matrix $\mathcal{K} = \left[|k(\epsilon_i(t), \epsilon'_j(t))|\right]$:

$$\hat{\sigma}_t \doteq \arg\max_{\sigma \in S(m)} k_t(\epsilon(t), \epsilon'(t); \sigma) = \arg\max_{\sigma \in S(m)} \sum_{i=1}^{m} \left|k(\epsilon_i(t), \epsilon'_{\sigma_i}(t))\right| \tag{48}$$

Similarly, we can define the determinant kernel k_d as:

$$k_d(\epsilon(t), \epsilon'(t); \sigma) = |\mathrm{det}U(\sigma)| = \prod_{i=1}^{m} \left|k\left(\epsilon_i(t), \epsilon'_{\sigma_i}(t)\right)\right| \tag{49}$$

which is a pointwise product of kernels and so is a kernel [40]. In this case the optimal matching $\hat{\sigma}_d$ is the solution to the assignment problem defined by the log-kernel matrix $\mathcal{K}_{log} = \left[\log\left|k(\epsilon_i(t), \epsilon'_j(t))\right|\right]$:

$$\hat{\sigma}_d \doteq \arg\max_{\sigma \in S(m)} k_d(\epsilon_i(t), \epsilon'_{\sigma_i}(t); \sigma) = \arg\max_{\sigma \in S(m)} \sum_{i=1}^{m} \log\left|k(\epsilon_i(t), \epsilon'_{\sigma_i}(t))\right| \tag{50}$$

The optimal matching problems $(48, 50)$ can be solved in $O(m^3)$ using the Hungarian algorithm [25]. We use these results to extend the kernels between linear systems $(34, 35)$ to include the effect of the input distributions. To do so, we apply the correlation matrix $U(\sigma)$ given in (46) in the calculation of the matrices for measurement noise $\Sigma[D, D]$ (31) and state noise $\Sigma[\{A, B, C\}, \{A', B', C'\}]$ (32). For trace kernels (34), we apply the correlation $U(\hat{\sigma}_t)$ corresponding to the optimal assignment $\hat{\sigma}_t$ solution to the additive matching problem (48). For determinant kernels (35) we apply $U(\hat{\sigma}_d)$ from the solution $\hat{\sigma}_d$ to the multiplicative assignment problem (50).

5 Experiments

5.1 Linearity and Gaussianity tests

We conducted simple tests for Gaussianity and linearity based on third-order temporal statistics [21]. Let cum_{3y} be the third-order cumulant of the stationary scalar process $y(t)$. The bispectrum S_{3y} is defined as the bidimensional Fourier transform of the cumulant: $S_{3y}(f_1, f_2) = \sum_{\tau_1=-\infty}^{\infty} \sum_{\tau_2=-\infty}^{\infty} \mathrm{cum}_{3y}(\tau_1, \tau_2) e^{-j2\pi f_1 \tau_1} e^{-j2\pi f_2 \tau_2}$. The bicoherence $\mathrm{bic}_{3y}(f_1, f_2)$ of $y(t)$ is defined as:

$$\mathrm{bic}_{3y}(f_1, f_2) = \frac{S_{3y}(f_1, f_2)}{\sqrt{S_{2y}(f_1 + f_2) S_{2y}(f_1) S_{2y}(f_2)}} \tag{51}$$

where is $S_{2y}(f)$ the power spectrum of $y(t)$.

If $y(t)$ is Gaussian, then its bispectrum is zero $S_{3y}(f_1, f_2) \equiv 0$ and the squared sample estimate of the bicoherence $|\hat{\mathrm{bic}}_{3y}(f_1, f_2)|^2$ is a central chi-squared random variable. A simple chi-squared test can then be devised to check for Gaussianity [21].

Additionally, the bicoherence can be used to test for linearity. If $y(t)$ is linear and non-Gaussian, i.e there exists some IID signal $n(t)$ such that (7) holds, then $\mathrm{bic}_{3y}(f_1, f_2)$ is a non-zero constant, and a test based on the comparison between sample and theoretical interquartile ranges of the corresponding chi-squared distribution can be derived [21].

We applied these tests to a dataset of 566 sequences of human gaits from the CMU motion

capture dataset [1]. Each motion component is treated separately as a scalar process, and in Fig. 1 we show the aggregated results, which can be interpreted as follows:

- $\mathrm{bic}_{3y}(f_1, f_2)$ is non-zero with near-one probability, hence human motion data are non-Gaussian processes.
- estimated and theoretical $\mathrm{bic}_{3y}(f_1, f_2)$ interquartile ranges are quite close, therefore the linearity hypothesis cannot be rejected.

This suggests modeling walking motions using a *non-Gaussian linear* process.

5.2 Linear Models for Synthesis

The goal of this first set of experiments is to validate the proposed models in synthesizing human motion. Given a motion sequence, we apply the learning algorithms of section 3 to estimate the parameters of our linear non-Gaussian models (5), and then use the inferred models to synthesize new motion. To demonstrate the flexibility of our approach we apply it to three representations of human motion:

- video: sequence of moving subject, images are centered and scaled on the moving body
- markers: 3D marker positions (from motion capture)
- angles: joint angles describing the pose of an articulated 3D human body model (from motion capture)

In these experiments we use two publicly available datasets, the CMU motion capture [1] and the CMU mobo dataset [20]. We show results for:

- video: fast walking sequence (340 frames), directory `moboJpg/04002/fastWalk/vr03_7` in the CMU Mobo dataset. The images have been scaled to 128×128 pixels and converted to 8-bit grayscale.
- markers: walking sequence `02_01.c3d` from CMU motion capture, consisting of 343 frames describing the 3D positions of 41 markers attached to the subject body.

- angles: sequence `02_01.amc` from CMU mocap (same as above), this time motion is represented as 52 joint angles of a skeletal model plus 6 DOFs for the reference frame.

The first step is to remove mean and linear trends from the original data. This is necessary given that our models represent zero-mean stationary processes. The best linear fit is removed from $3D$ position data, while in body joint angles and image data no significant linear trend is observed and so only the mean is subtracted. For synthesis, mean and linear trends are added back to the output sequence produced by the learned model to give the final motion.

A second preprocessing step consists in applying principal component analysis (PCA) to reduce the dimensionality of the observations. This step is not required theoretically but necessary in practice since we are dealing with short realizations (few hundred frames) of high-dimensional processes. For motion capture data, we verified experimentally that 8 PCA bases are sufficient to synthesize sequences that are perceptually indistinguishable from the original. For image data, the required dimensionality is higher, in the sequence shown here we use 16 components.

Given a low-dimensional motion sequence $y(t)$, we use the subspace identification algorithm of section 3.1 to estimate the set of models $M_l = \{A, K_l, C, R_l\}$ $l = 1, \cdots, L$ representing the second-order statistics of the observed process. For estimating the state sequence (13) we adopted the algorithm described on page 131 of [34]. We determined experimentally the dimensions (n_d, n_s) of the deterministic and stochastic components of the process to be $(8, 8)$, $(6, 6)$ and $(6, 6)$ respectively for the video, marker and angle motion data sequences. Then using the closed-form equations derived in 3.1 in the appendix we obtain the matrice estimates $\hat{A}_d, \hat{A}_s, \hat{C}_s, \hat{C}_d, G_s$ and $\hat{\Lambda}_s$. By solving the Riccati equation [5] we obtain the set of input-related matrices K_l, R_l generating the same second-order statistics of $y(t)$. In our experiments we noticed that zeros with very small norm undermine the robustness of the subsequent deconvolution step. We overcame this problem by introducing a threshold on the minimum norm of reflected zeros, here set to 0.1. The threshold helps also pruning the number of combinations to be considered, tipically giving 2- to 8-fold reductions.

For each candidate model $M_l = \{A, K_l, C, R_l\}$, we use its inverse M_l^{-1} to deconvolve the process $y(t)$ and recover the white input $n_l(t)$, and the initial state x_0. Then for each input sequence we compute the third-order temporal independence score $\rho(n_l)$ (21) (setting $N = 10$) and select the model $M_{\hat{l}}$ that provides maximum independence: $\hat{l} = \arg\max_l \rho(n_l)$. In Fig. 2 we show the independence scores for the candidate models computed from the three test sequences. We see that the minimum phase model is never the most temporally independent, thus providing further evidence that human dynamics is not minimum phase.

The last inference step is the estimation of the mixing matrix D and the input component distributions q_i. As described in section 3.3, D is estimated using a standard ICA algorithm [6], while the input distributions q_i are represented by sample histograms. Fig. 3 shows some component distributions estimated from the data sequences.

Once we have the parameters $\{\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{x}_0\}$ of the linear model (3) and the distributions q_i of the input components ϵ_i, we can generate a new sequence with the same temporal and spatial statistics of the original motion. We start by setting $x(0) = \hat{x}_0$. Then for $t = 1, \cdots, T$:

- form $\epsilon(t)$ by independently drawing $\epsilon_i(t) \sim q_i(\epsilon_i)$
- apply the input $\epsilon(t)$ to the linear system (3) to obtain the synthetic output $y(t)$
- apply the PCA basis to lift the generated signal $y(t)$ to the measurement space.
- add mean and linear trends back to obtain the final motion.

In Fig. 4 we show some sample frames of synthesis for motion capture data (both marker positions and joint angles) from minimum-phase and optimal non-minimum phase systems. As expected, motion produced by the non-minimum phase model is perceptually closer to the original.

For video sequences, the low quality of the synthetic images due to the linear PCA projection does not allow to visually appreciate the difference between outputs of minimum and non-minimum phase models. However, as we show in Fig. 5, correctly representing periodic modes and non-Gaussian statistics allows to largely outperform the results obtained with standard Gaussian linear models such as "dynamic textures" [16].

5.3 Kernels for Gait Recognition

In this section we present results on the applications of the proposed kernels for non-Gaussian systems (34) to the problem of classifying human gaits. These experiments are based on the CMU Mobo dataset [20]. The goal is to identify the 4 classes of walking motions (normal walk, fast walk, walk with ball and walk on inclined treadmill) performed by the 24 subjects in the dataset. We use only the sequences taken from the same viewpoint (camera $vr03\_7$). Each sequence is 340 frames long.

Directly modeling gait video sequences using linear systems of the form (5) is a viable approach to synthesis but does not yield satisfactory results for recognition problems. It is necessary to derive a representation insensitive to changes in background, illumination and appearance of the subject performing the motion, and a simple choice are binary silhouettes of the moving body. We used the ones provided in the Mobo dataset, obtained by simple differencing with a background image followed by thresholding. Given that the extracted silhouettes are rather noisy (in particular in the inclined walk sequences large parts of the treadmill are labeled as foreground), we derive a set of coarse features providing a robust description of the shape. After evaluating several alternatives including PCA projection and Hu moments, we found that the projection features proposed in [15] are robust and effective representations for human silhouettes.

In Fig. 6 we show a sample image from background subtraction and the corresponding representation with the projection features. Given a binary silhouette, the projection features encode the distance of the points on the silhouette from lines passing through its center of mass. The bounding box of the silhouette is divided uniformly in $2n$ region, n to each side of the projection line, and for each region the average distance from the line is computed. In our experiments we used 2 lines (horizontal and vertical) and $n = 8$ features on both side, for a total 32 components (Fig. 6).

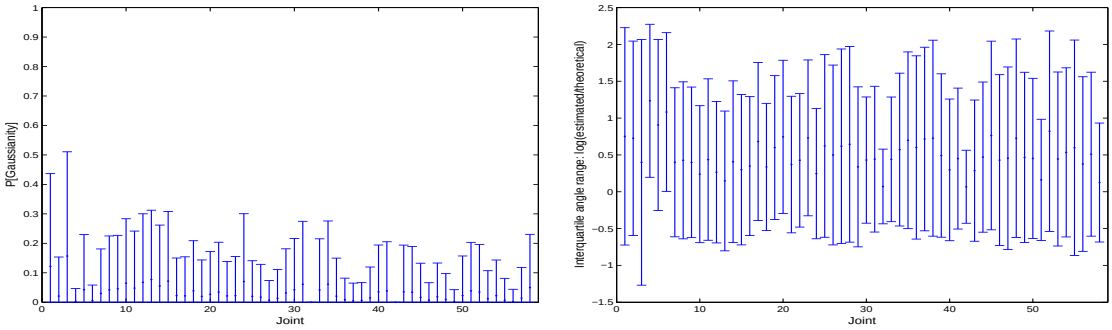
On the feature trajectories extracted from a video sequence, we apply the proposed learning algorithm to estimate the parameters of the linear non-Gaussian model (5). As before, in order to obtain better estimates it is advisable to reduce the dimensionality of the data by

PCA projection, here we use $m = 8$ components. The parameters of the learned models are $n_d = 8$ components for the deterministic and $n_s = 4$ components for the stochastic part. We observed that in this case the effect of phase is marginal. This may be due the coarseness of the representation, which masks the fine discriminative power of the higher-order *temporal* statistics.

Once a set of model parameters $\{A, B, C, D, x_0, q_1, \cdots, q_m\}$ are estimated from each sequence in the dataset, we can apply the kernels (34, 35) to measure similarity between models. In these experiments we used the trace kernels, which offer computational advantages over determinant kernels and allow to separate the effects of the stochastic and periodic components.

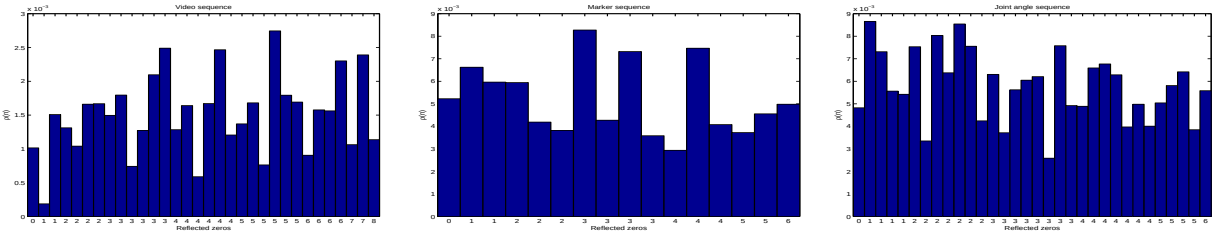
First we investigate the effects of the proposed alignment of initial conditions on the performances of kernel in matching gait processes. Aligning the initial states proves to be particularly effective when the correction is restricted to the periodic part of the model $\{A_d, C_d, x_{0,d}\}$. In Fig. 7 we compare the standard trace kernel on the initial condition correlation (30) (as in [44]) to the aligned kernel as defined in (37), with maximum delay $T = 20$ (38). We can see how the aligned kernel provides a similarity measure which is insensitive to delays, while the standard kernel exhibits a periodic behaviour.

For each learned model pair in the dataset we then proceed to compute the full trace kernels (34). These are made of two terms: The similarity between the deterministic part of the systems encoded in periodic components and aligned initial states (37), and the matching between the stochastic parts, represented by kernels between input statistics (31, 32). In Fig. 8 we plot the confusion matrices showing the distances (36) between learned models defined by initial state trace kernels (left) and the full trace kernels, including input distributions (right). It is evident that the inclusion of the stochastic part modeled by the input statistics improves the gait discrimination performances, visible by the block diagonal structure of the corresponding confusion matrix and the higher number of same-gait nearest neighbor matches.



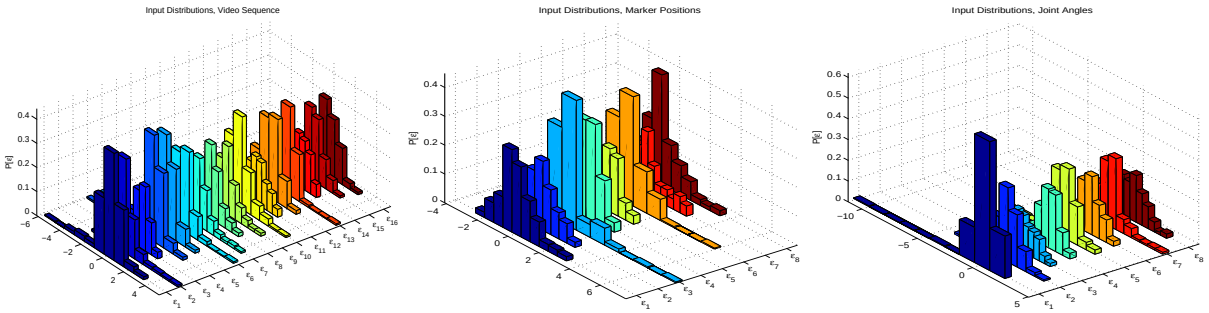
(a) (b)

Figure 1: Linearity and Gaussianity tests on 566 sequences of walking, running, jumping, hopping, climbing and limping from the CMU motion capture dataset. The motion is represented with 58 degrees of freedom (DOFs): 3 coordinates for global position, 3 angles for global orientation, and 52 joint angles describing the body pose. The scalar test [21] is conducted independently once on each motion data component of each sequence, here for each DOF we show mean and standard deviation of the results in error bar plots. (a) Probability that the bispectrum of the observed sequence is zero. Since this probability is very low, we can conclude that the bispectrum is non-zero and thus human motion data is non-Gaussian. (b) Ratio of theoretical (assuming linearity hence constant bichoerence) vs. sample interquartile bicoherence ranges (in log scale). The distribution around zero of the log-ratio suggests that the linearity hypothesis cannot be ruled out.



(a) (b) (c)

Figure 2: Temporal dependence for human motion input processes. We show the independence score (21) of the deconvolved white input as vertical bars, one for each possible realization (minimum and non-minimum phase) of the second-order statistics of the data. The values are sorted by number of unstable zeros of the corresponding model, reported below each bar. We see how the minimum phase input is never the mostly temporal independent. Results for: (a) video sequence of fast walking from the Mobo dataset, having $n_s = 8$ and $L = 32$ candidate models; (b) marker sequence, $n_s = 6$ and $L = 16$; (c) joint angle sequence, $n_s = 6$ and $L = 32$.



(a) (b) (c)

Figure 3: Input component distributions. We show the sample histograms computed on the components of the IID non-Gaussian input $\epsilon(t)$ to the non-minimum phase system estimated from three human motion sequences (see Fig. 2): (a) walking sequence from the Mobo dataset, (b) marker sequence and (c) joint angle sequence. From these plots we can see that the estimated input distributions have long asymmetrical tails and thus are non-Gaussian.

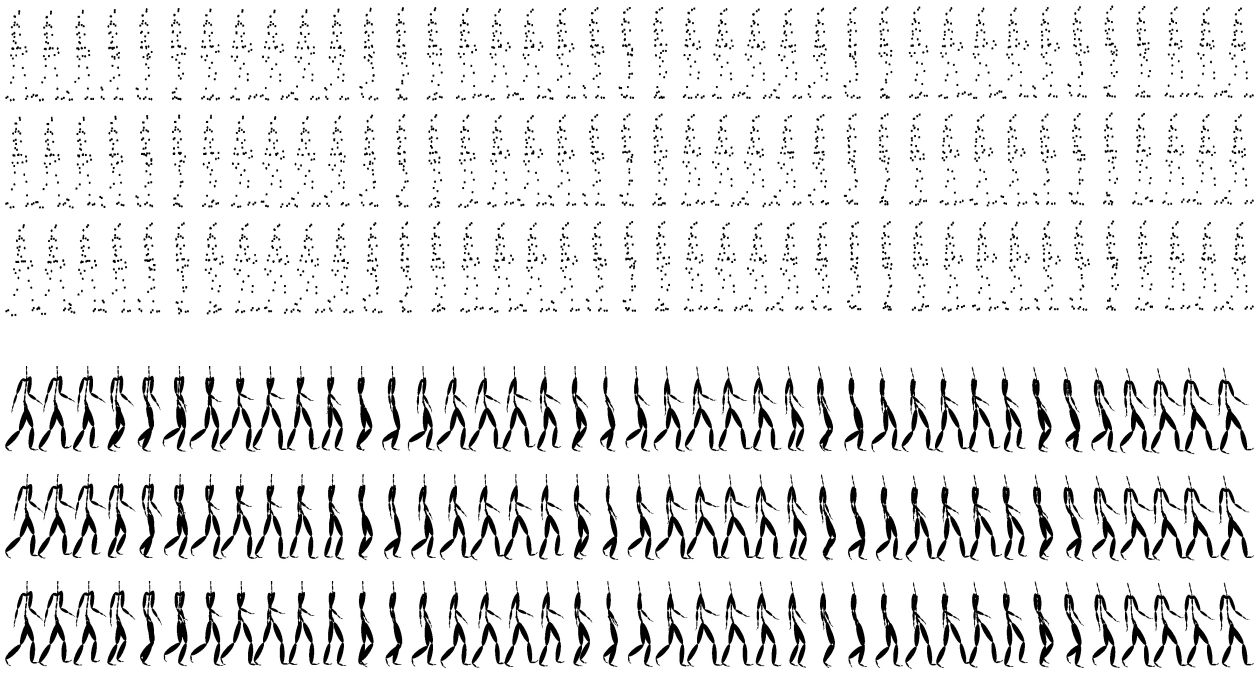


Figure 4: Synthesis results for models learned from motion capture data, both for marker positions (first three rows) and joint angle representations (last three rows). For each group, the first row shows the original sequence, the second row displays the synthesis obtained with the minimum phase system, while the last row shows the synthesis from the optimal non-minimum phase system. It is perceivable (see also the movies, downloadable from `http://www.cs.ucla.edu/~bissacco/dynamicICA`) the better fidelity of the non-minimum phase synthesis to the character of the original motion.

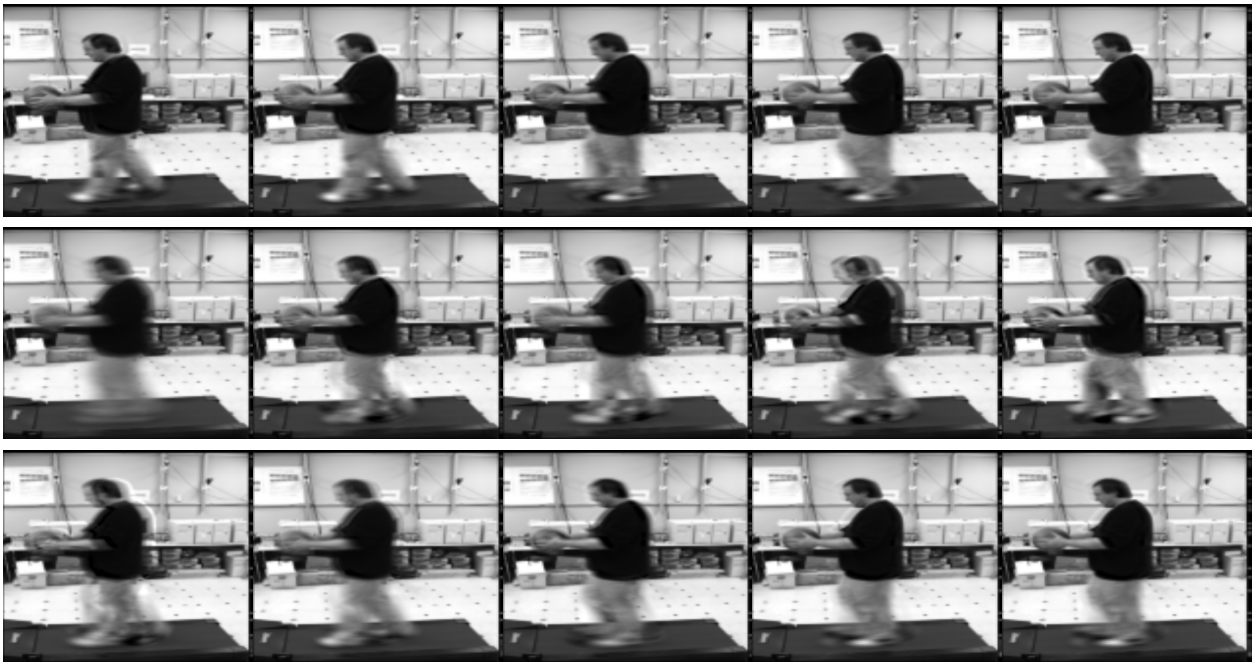


Figure 5: Comparison between our non-Gaussian linear models and standard Gaussian ARMA models (dynamic textures [16]). We set the dimension of the state in both systems to 16, assigning in our model $n_d = 8$ components to the periodic part and $n_s = 8$ components to the stochastic part. First row shows original sequence after PCA projection, second row displays corresponding frames produced by a dynamic texture model, last row is output of our model. These few frames are sufficient to show the ability of our model to deliver better quality in the synthesis (notice how images are less blurry and more synchronized with the original compared to dynamic textures), thus validating the importance of explicitly representing periodic modes and high-order statistics. See also the entire movies at `http://www.cs.ucla.edu/~bissacco/dynamICA`.

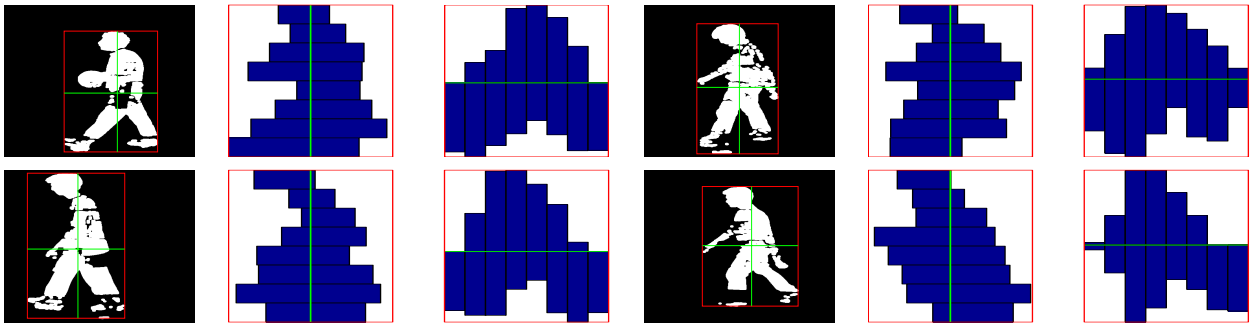


Figure 6: Sample silhouettes and associated shape features. First and fourth columns shows some sample background subtraction frames from the gait dataset [20]: walking with ball, normal walk, fast walk and inclined walk. Superimposed to the binary silhouette we plot the bounding box (red) and the horizontal and vertical lines passing through the center of mass used to extract the features. On columns $(2, 5)$ and $(3, 6)$ we show the features obtained by computing the distance of the points on the two sides of the silhouette to respectively the vertical and horizontal lines, discretized to $n_f = 8$ values.

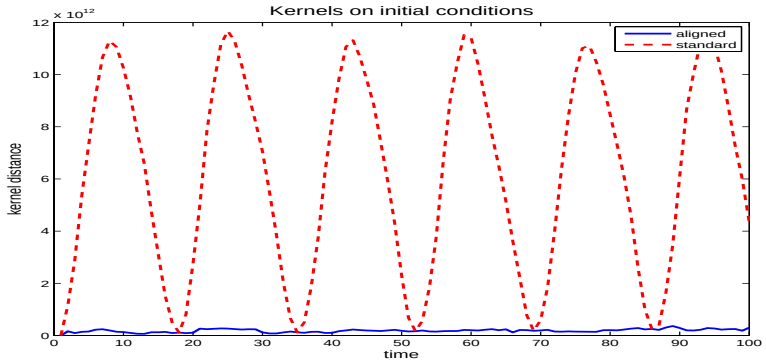


Figure 7: Standard vs. aligned kernels on initial conditions. Here we show the kernel distances between a model learned from a segment of a walking sequence and the set of models learned form the following segments of the same sequence. The x axis denotes the time delay between the two sequences, while on y we plot the kernel distances, both for the standard (30) (dashed line) and aligned (37) (solid line) kernels, where the latter are computed with maximum delay $T = 20$ (38). We see how periodic bias present in the standard kernel practically disappears when using the aligned kernel.

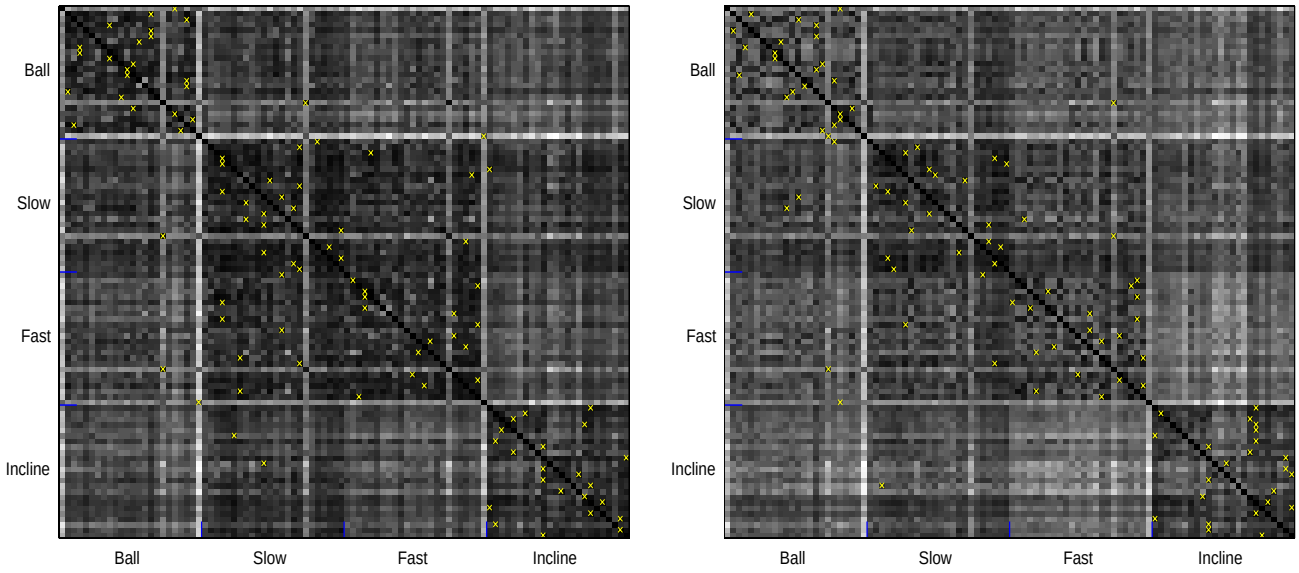


Figure 8: State and input kernel distances. We show the confusion matrices representing trace kernel distances between non-Gaussian linear models learned from walking sequences in the Mobo dataset. There are 4 motion classes and 24 individuals performing these motions, for a total of 96 sequences. For each sequence we learn a linear model (5) and then measure distance between models by the trace kernels. On the left we show results using kernels on initial states only, on the right we display the confusion matrix obtained from the trace kernels that include the effect of the input (34). For each row a cross indicates the nearest neighbor. It is clear how the additional information provided by the input statistics results in improved gait classification performances: we have 17 (17.7%) nearest neighbors mismatches (i.e. closest models that do not belong to the same gait class) using the state-only distance, while only 9 (9.3%) with the complete trace kernel distance.

References

[1] Cmu. carnegie-mellon mocap database, 2003. http://mocap.cs.cmu.edu.

[2] D. Bauer. Asymptotic properties of subspace estimators. *Automatica*, 41:359–376, 2005.

[3] D. Bauer and M. Wagner. Estimating cointegrated systems using subspace algorithms. *Journal of Econometrics*, 111:47–84, 2002.

[4] P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, Vol. 9, No. 6:1196–1217, 1981.

[5] A. Bissacco, A. Chiuso, and S. Soatto. Classification and recognition of dynamical models. Technical report, Computer Science Department, University of California - Los Angeles, CA, 2006.

[6] T. Blaschke and L. Wiskott. Cubica: Independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. *IEEE Transactions on Signal Processing*, 52:1250–1256, 2004.

[7] M. Boumahdi. Blind identification using the kurtosis with applications to field data. *Signal Processing*, 48:205–216, 1996.

[8] R. Brockett. *Finite Dimensional Linear Systems*. John Wiley and Sons, Inc., 1970.

[9] A. Chiuso. Asymptotic variance of closed-loop subspace identification algorithms. *IEEE Trans. on Aut. Control, to appear*, 2006. available at http://www.dei.unipd.it/∼chiuso.

[10] A. Chiuso and G. Picci. Subspace identification by orthogonal decomposition. In *Proc. 14th IFAC World Congress*, volume I, pages 241–246, 1999.

[11] A. Chiuso and G. Picci. On the ill-conditioning of subspace identification with inputs. *Automatica*, 40(4):pp. 575–589, 2004.

[12] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. Wiley, 2003.

[13] K. De Coch and B. De Moor. Subspace angles and distances between arma models. *Proc. of the Intl. Symp. of Math. Theory of Networks and Systems*, 2000.

[14] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36:287–314, 1994.

[15] F. Cuzzolin. Using bilinear models for view-invariant action and identity recognition. *Proc. CVPR*, 2006.

[16] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51 (2):91–109, 2003.

[17] A. Eriksson, P. Stoica, and T. Soderstrom. Markov-based eigenanalysis method for frequency estimation. *IEEE Trans. on Signal Processsing*, Vol. 42, No. 3:586–594, 1994.

[18] I. Goethals, K. Pelckmans, J. A. K. Suykens, and B. D. Moor. Subspace identification of hammerstein systems using least squares support vector machines. *IEEE Trans. on Automatic Control*, Vol. 50, No. 10:1509–1519, Oct. 2005.

[19] G.H. Golub and C.R. Van Loan. *Matrix Computation*. The Johns Hopkins Univ. Press., 1989.

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

[21] M. J. Hinich. Testing for gaussianity and linearity of a stationary time series. *J. Time Series Analysis*, Vol 3:169–76, 1982.

[22] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[23] T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, 2000.

[24] M. Kristensson, M. Jansson, and B. Ottersten. Further results and insights on subspace based sinusoidal frequency estimation. *IEEE Trans. on Signal Processing*, Vol. 49, No. 12:2962–2974, 2001.

[25] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83, 1955.

[26] J. Levine. Finite dimensional filters for a class of nonlinear systems and immersion in a linear system. *SIAM J. Control and Optimization*, 25(6):1430–1439, 1987.

[27] A. Lindquist and G. Picci. A geometric approach to modelling and estimation of linear stochastic systems. *Journal of Mathematical Systems, Estimation and Control*, 1:241–333, 1991.

[28] A. Lindquist and G. Picci. Geometric methods for state-space identification. In S. Bittanti and G. Picci, editors, *Identification, Adaptation, Learning*, pages 1–69. Springer Verlag, 1996.

[29] L. Ljung. *System Identification; Theory for the User*. Prentice Hall, 1997.

[30] C. L. Mallows. A note on asymptotic joint normality. *Ann. of Mathematical Statistics*, 43:508–515, 1972.

[31] R. Martin. A metric for arma processes. *IEEE Trans. on Signal Processing*, 48(4):1164–1170, 2000.

[32] J. Mourjopoulos, P. M. Clarkson, and J. K. Hammond. A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signal. *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1858–1861, 1982.

[33] P. Mullhaupt, B. Srinivasan, and D. Bonvin. On the nonminimum-phase characteristics of two-link underactuated mechanical systems. *In Proc. Conference on Decision and Control*, 1998.

[34] P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer, Norwell, 1996.

[35] A. Papoulis. Predictable processes and wold's decomposition: an review. *IEEE Transactions on Acoustic, Speech and Signal Processing*, 33(4):933–938, 1985.

[36] G. Picci and T. Katayama. Stochastic realization with exogenous inputs and "subspace methods" identification. *Signal Processing*, 52:145–160, 1996.

[37] R. H. Roy. Esprit - a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-34:1340–1342, Oct. 1986.

[38] Y.A. Rozanov. *Stationary Random Processes*. Holden-Day, San Francisco, 1967.

[39] R. O. Schmidt. *A signal subspace approach to multiple emitter location and spectral estimation*. PhD thesis, Stanford University, Nov. 1981.

[40] B. Schoelkopf and A. J. Smola. *Learning with Kernels*. The MIT Press, 2002.

[41] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, 1989.

[42] A. Swami, G. Giannakis, and S. Shamsunder. Multichannel arma processes. *IEEE Trans. on Signal Processing*, Vol. 42, No. 4:898–913, 1994.

[43] P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993.

[44] 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.