Chapter 8
Distances on Spaces of High-Dimensional Linear Stochastic Processes: A Survey

Bijan Afsari and René Vidal

Abstract In this paper we study the geometrization of certain spaces of stochastic processes. Our main motivation comes from the problem of pattern recognition in high-dimensional time-series data (e.g., video sequence classification and clustering). In the first part of the paper, we provide a rather extensive review of some existing approaches to defining distances on spaces of stochastic processes. The majority of these distances are, in one or another, based on comparing power spectral densities of the processes. In the second part, we focus on the space of processes generated by (stochastic) linear dynamical systems (LDSs) of fixed size and order, on which we recently introduced a class of group action induced distances called the alignment distances. This space is a natural choice in some pattern recognition applications and is also of great interest in control theory. We review and elaborate on the notion of alignment distance. Often it is convenient to represent LDSs in state-space form, in which case the space (more precisely manifold) of LDSs can be considered as the base space of a principal fiber bundle comprised of state-space realizations. This is due to a Lie group action symmetry present in the state-space representation of LDSs. The basic idea behind the alignment distance is to compare two LDSs by first aligning a pair of their realizations along the respective fibers. Upon a standardization (or bundle reduction) step this alignment process can be expressed as a minimization problem over orthogonal matrices, which can be solved efficiently. The alignment distance differs from most existing distances in that it is a structural or generative distance, since in some sense it compares how two processes are generated. We also briefly discuss averaging LDSs using the alignment distance via minimizing a sum of the squares of distances (namely, the so-called Fréchet mean).

B. Afsari (✉) · R. Vidal
Center for Imaging Science, Johns Hopkins University, Baltimore MD 21218, USA
e-mail: bijan@cis.jhu.edu

R. Vidal
e-mail: rvidal@cis.jhu.edu

F. Nielsen (ed.), *Geometric Theory of Information*,
Signals and Communication Technology, DOI: 10.1007/978-3-319-05317-2_8,

318735_1_En_8_Chapter ✓ TYPESET ☐ DISK ☐ LE ✓ CP Disp.:27/3/2014 Pages: 24 Layout: T1-Standard

Keywords Stochastic processes · Pattern recognition · Linear dynamical systems · Extrinsic and intrinsic geometries · Principal fiber bundle · Generalized dynamic factor model · Minimum phase · Spectral factorization · All-pass filter · Hellinger distance · Itakura-Saito divergence · Fréchet mean

8.1 Introduction and Motivation

Pattern recognition (e.g., classification and clustering) of time-series data is important in many real world data analysis problems. Early applications include the analysis of one-dimensional data such as speech and seismic signals (see, e.g., [48] for a review). More recently, applications in the analysis of video data (e.g., activity recognition [1]), robotic surgery data (e.g., surgical skill assessment [12]), or biomedical data (e.g., analysis of multichannel EEG signals) have motivated the development of statistical techniques for the analysis of high-dimensional (or vectorial) time-series data.

The problem of pattern recognition for time-series data, in its full generality, needs tools from the theory of statistics on stochastic processes or function spaces. Thus it bears relations with the general problem of inference on (infinite dimensional) spaces of stochastic processes which is a quite sophisticated mathematical theory [30, 59]. However, at the same time, the pattern recognition problem is more complicated since, in general, it involves not only inference but also learning. Learning and inference on infinite dimensional spaces obviously can be daunting tasks. In practice, there have been different grand strategies proposed in dealing with this problem (e.g., see [48] for a review). In certain cases it is reasonable and advantageous from both theoretical and computational points of view to simplify the problem by assuming that the observed processes are generated by models from a specific finite-dimensional class of models. In other words, one could follow a parametric approach based on *modeling* the observed time series and then performing statistical analysis and inference on a finite dimensional *space of models* (instead of the space of the observed *raw* data). In fact, in many real-world instances (e.g., video sequences [1, 12, 22, 60] or econometrics [7, 20, 24]), one could model the observed high-dimensional time series with low-order Linear Dynamical Systems (LDSs). In such instances the mentioned strategy could prove beneficial, e.g., in terms of implementation (due to significant compression achieved in high dimensions), statistical inference, and synthesis of time series. For 1-dimensional time-series data the success of Linear Predictive Coding (i.e., auto-regressive (AR) modeling) modeling and its derivatives in modeling speech signals is a paramount example [26, 49, 58]. These motivations lead us to state the following prototype problem:

Problem 1 (*Statistical analysis on spaces of LDSs*) Let $\{\boldsymbol{y}^i\}_{i=1}^N$ be a collection of p-dimensional time series indexed by time t. Assume that each time series $\boldsymbol{y}^i = \{\boldsymbol{y}_t^i\}_{t=1}^\infty$ can be approximately modeled by a (stochastic) LDS M_i of output-input size (p, m) and order n[1] realized as

[1] Typically in video analysis: $p \approx 1000$–10000, $m, n \approx 10$ (see e.g., [1, 12, 60]).

318735_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:27/3/2014 Pages: 24 Layout: T1-Standard

$$
\begin{aligned}
\boldsymbol{x}_t^i &= A_i \boldsymbol{x}_{t-1}^i + B_i \boldsymbol{v}_t,\\
\boldsymbol{y}_t^i &= C_i \boldsymbol{x}_t^i + D_i \boldsymbol{v}_t, \quad (A_i, B_i, C_i, D_i) \in \widetilde{\mathcal{SL}}_{m,n,p} = \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n} \times \mathbb{R}^{p\times m}
\end{aligned}
\tag{8.1}
$$

where $\boldsymbol{v}_t$ is a common stimulus process (e.g., white Gaussian noise with identity covariance)[2] and where the realization $R_i = (A_i, B_i, C_i, D_i)$ is learnt and assumed to be known. The problem is to: (1) Choose an appropriate space $\mathcal{S}$ of LDSs containing the learnt models $\{M_i\}_{i=1}^N$, (2) geometrize $\mathcal{S}$, i.e., equip it with an appropriate geometry (e.g., define a distance on $\mathcal{S}$), (3) develop tools (e.g., probability distributions, averages or means, variance, PCA) to perform statistical analysis (e.g., classification and clustering) in a computationally efficient manner.

The first question to ask is why to model the processes using the state-space model (representation) (8.1)? Recall that processes have equivalent ARMA and state-space representations and model (8.1) is quite general and with n large enough it can approximate a large class of processes. More importantly, state-space representations (especially in high dimensions) are often more suitable for parameter learning or system identification. In important practical cases of interest such models conveniently yield more parsimonious parametrization than vectorial ARMA models , which suffer from the *curse of dimensionality* [24]. The curse of dimensionality in ARMA models stems from the fact that for p-dimensional time series if p is very large the number of parameters of an ARMA model is roughly proportional to p^2, which could be much larger than the number of data samples available pT, where T is the observation time period (note that the autoregressive coefficient matrices are very large $p \times p$ matrices). However, in many situations encountered in real world, state-space models are more effective in overcoming the curse of dimensionality [20, 24]. The intuitive reason, as already alluded to, is that often (very) high-dimensional time series can be well approximated as being generated by a *low order* but high-dimensional dynamical system (which implies *small* n despite *large* p in the model (8.1)). This can be attributed to the fact that the components of the observed time series exhibit correlations (cross sectional correlation). Moreover, the contaminating noises also show correlation across different components (see [20, 24] for examples of exact and detailed assumptions and conditions to formalize these intuitive facts). Therefore, overall the number of parameters in the state-space model is small compared with p^2 and this is readily reflected in (or encoded by) the small size of the dynamics matrix A_i and the thinness of the observation matrix C_i in (8.1).[3]

[2] Note that in a different or more general setting the noise at the output could be a process $\boldsymbol{w}_t$ different (independent) from the input noise $\boldsymbol{v}_t$. This does not cause major changes in our developments. Since the output noise usually represents a perturbation which *cannot* be modeled, as far as Problem 1 is concerned one could usually assume that $D_i = 0$.

[3] Note that we are not implying that ARMA models are incapable of modeling such time series. Rather the issue is that general or unrestricted ARMA models suffer from the curse of dimensionality in the identification problem, and the parametrization of a restricted class of ARMA models with small number of parameters is complicated [20]. However, at the same time, using state-space models it is easy to overcome the curse of dimensionality and this approach naturally leads to simple and effective identification algorithms [20, 22].

Also, in general, state-space models are more convenient for computational purposes than vectorial ARMA models. For example, in the case of high-dimensional time series most effective estimation methods are based on state-space domain system identification rooted in control theory [7, 41, 51]. Nevertheless, it should be noted that, in general, the identification of multi-input multi-output (MIMO) systems is a subtle problem (see Sect. 8.4 and e.g., [11, 31, 32]). However, for the case where $p > n$, there are efficient system identification algorithms available for finding the state-space parameters [20, 22].

Notice that in Problem 1 we are assuming that all the LDSs have the same order n (more precisely the minimal order, see Sect. 8.3.3.1). Such an assumption might seem rather restrictive and a more realistic assumption might be all systems having order not larger than n (see Sect. 8.5.1). Note that since in practice real data can be only *approximately* modeled by an LDS of fixed order, if n is not chosen too large, then gross over-fitting of n is less likely to happen. From a practical point of view (e.g., implementation) fixing the order for all systems results in great simplification in implementation. Moreover, in classification or clustering problems one might need to combine (e.g., average) such LDSs for the goal of replacing a class of LDSs with a representative LDS. Ideally one would like to define an average in a such a way that LDSs of the same order have an average of the same order and not higher, otherwise the problem can become intractable. In fact, most existing approaches tend to dramatically increase the order of the average LDS which is certainly undesirable. Therefore, intuitively, we would like to consider a space $\mathcal{S}$ in which the order of LDSs is fixed or limited. From a theoretical point of view also this assumption allows us to work with nicer mathematical spaces namely smooth manifolds (see Sect. 8.4).

Amongst the most widely used classification and clustering algorithms for static data are the k-nearest neighborhood and k-means algorithms, both of which rely on a notion of distance (in a feature space) [21]. These algorithms enjoy certain universality properties with respect to the probability distributions of the data; and hence in many practical situations where one has little prior knowledge about the nature of the data they prove to be very effective [21, 35]. In view of this fact, in this paper we focus on the notion of distance between LDSs and the stochastic processes they generate. Hence, a natural question next is what space and what type of distance on it? In Problem 1, obviously, the first two steps (which are the focus of this paper) have significant impacts on the third one. One has different choices for the space $\mathcal{S}$, as well as, for geometries on that space. The gamut ranges from an *infinite dimensional linear* space to a *finite dimensional (non-Euclidean) manifold*, and the geometry can be either *intrinsic* or *extrinsic* . By an intrinsic geometry we mean one in which a shortest path between two points in a space stays in the space, and by an extrinsic geometry we mean one where the distance between the two points is measured in an *ambient* space. In the second part of this paper, we study our recently developed approach which is somewhere in between: to design an *easy-to-compute* extrinsic distance, while keeping the ambient space *not* too large.

This paper is organized as follows: In Sect. 8.2, we review some existing approaches in geometrization of spaces of stochastic processes. In Sect. 8.3, we focus on processes generated by LDSs of fixed order, and in Sect. 8.4, we study smooth

318735_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:27/3/2014 Pages: 24 Layout: T1-Standard

fiber bundle structures over spaces of LDSs generating such processes. Finally, in Sect. 8.5, we introduce our class of group action induced distances namely the *alignment* distances. The paper is concluded in Sect. 8.6. To avoid certain technicalities and just to convey the main ideas the proofs are omitted and will appear elsewhere. We should stress that the theory of alignment distances on spaces of LDSs is still under development; however, its basics have appeared in earlier papers [1–3]. This paper for most parts is an extended version of [3].

8.2 A Review of Existing Approaches to Geometrization of the Spaces of Stochastic Processes

This review, in particular, since the subject appears in a range of disciplines is non-exhaustive. Our emphasis is on the core ideas in defining distances on spaces of stochastic processes rather than enumerating all such distances. Other sources to consult may include [9, 10, 25]. In view of Problem 1, our main interest is in the finite dimensional spaces of LDSs of fixed order and the processes they generate. However, since such a space can be embedded in the larger infinite dimensional space of "virtually all processes," first we consider the latter.

Remark 1 We shall discuss several distance-like measures some of which are known as distance in the literature. We will try to use the term *distance* exclusively for a true distance namely one which is symmetric, positive definite and obeys the triangle inequality. Due to convention or convenience, we still may use the term distance for something which is not a true distance, but the context will be clear. A distance-like measures is called a divergence it is only positive definite and it is called pseudo-distance, if it is symmetric and obeys the triangle inequality but it is only positive semi-definitive (i.e., a zero distance between two processes does not imply that they are the same). As mentioned above, our review is mainly to show different schools of thought and theoretical approaches in defining distances. Obviously, when it comes to comparing these distances and their effectiveness (e.g., in terms of recognition rate in a pattern recognition problem) ultimately things very much depend on the specific application at hand. Although we should mention that for certain 1D spectral distances there has been some research about their relative discriminative properties; especially for applications in speech processing, the relation between of the distances to the human auditory perception system has been studied (see e.g., [9, 25, 26, 29, 49, 54]). Perhaps one aspect that one can judge about rather comfortably and independently of the specific problem is the associated computational costs of calculating the distance and other related calculations (e.g., calculating a notion of average). In that regard, for Problem 1, when the time-series dimension p is very large (e.g., in video classification problems) our introduced alignment distance (see Sect. 8.5) is cheaper to calculate relative to most other distances and also renders itself quite effective in defining a notion of average [1].

318735_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:27/3/2014 Pages: 24 Layout: T1-Standard

Remark 2 Throughout the paper, unless otherwise stated, by a *process* we mean a (real-valued) discrete-time wide-sense (or second order) stationary zero mean Gaussian regular stochastic process (i.e., one with no deterministic component). Some of the language used in this paper is borrowed from the statistical signal processing and control literature for which standard references include [40, 56]. Since we use the Fourier and z-transforms often and there are some disparities between the definitions (or notations) in the literature we review some terminologies and establish some notations. The z-transform of a matrix sequence $\{\mathbf{h}_t\}_{-\infty}^{+\infty}(\mathbf{h}_t \in \mathbb{R}^{p\times m})$ is defined as $H(z) = \sum_{-\infty}^{+\infty} \mathbf{h}_t z^{-t}$ for z in the complex plane $\mathbb{C}$. By evaluating $H(z)$ on the unit circle in the complex plane $\mathbb{C}$ (i.e., by setting $z = e^{i\omega}, \omega \in [0, 2\pi]$) we get $H(e^{i\omega})$, the Fourier transform of $\{\mathbf{h}_t\}_{-\infty}^{+\infty}$, which sometimes we denote by $H(\omega)$. Note that the z-transform of $\{\mathbf{h}_{-t}\}_{-\infty}^{+\infty}$ is $H(z^{-1})$ and its Fourier transform is $H(e^{-i\omega})$, and since we deal with real sequences it is the same as $\overline{H(e^{i\omega})}$, the complex conjugate of $H(e^{i\omega})$. Also any matrix sequence $\{\mathbf{h}_t\}_0^{+\infty}$ defines (causal) a linear filter via the convolution operation $\boldsymbol{y}_t = \sum_{\tau=0}^{\infty} h_\tau \boldsymbol{\epsilon}_{t-\tau}$ on the m-dimensional sequence $\boldsymbol{\epsilon}_t$. In this case, we call $H(\omega)$ or $H(z)$ the *transfer function* of the *filter* and $\{\mathbf{h}_t\}_0^{+\infty}$ the impulse response of the filter. We also say that $\boldsymbol{\epsilon}_t$ is filtered by H to generate $\boldsymbol{y}_t$. If $H(z)$ is an analytic function of z outside the unit disk in the complex plane, then filter is called asymptotically stable. If the transfer function $H(z)$ is a *rational* matrix function of z (meaning each entry of $H(z)$ is a rational function of z), then the filter has a *finite* order state-space (LDS) realization in the form (8.1). The smallest (*minimal*) order of such an LDS can be determined as the sum of the orders of the denominator polynomials (in z) in the entries appearing in a specific representation (factorization) of $H(z)$, known as the *Smith-McMillan* form [40]. For a square transfer function this number (known as the *McMillan degree*) is, generically, equal to the order of the denominator polynomial in the determinant of $H(z)$. The roots of these denominators are the eigenvalues of the A matrix in the minimal state-space realization of $H(z)$ and the system is asymptotically stable if all these eigenvalues are inside the unit disk in $\mathbb{C}$.

8.2.1 Geometrizing the Space of Power Spectral Densities

A p-dimensional process $\boldsymbol{y}_t$ can be *identified* with its $p \times p$ *covariance sequence* sequences $C_{\boldsymbol{y}}(\tau) = \mathbb{E}\{\boldsymbol{y}_t \boldsymbol{y}_{t-\tau}^\top\}$ $(\tau \in \mathbb{Z})$, where $^\top$ denotes matrix transpose and $\mathbb{E}\{\cdot\}$ denotes the expectation operation under the associated probability measure. Equivalently, the process can be identified by the Fourier (or z) transform of its covariance sequence, namely the *power spectral density* (PSD) $P_{\boldsymbol{y}}(\omega)$, which is a $p \times p$ Hermitian positive semi-definite matrix for every $\omega \in [0, 2\pi]$.[4] We denote the space of all $p \times p$ PSD matrices by $\mathcal{P}_p$ and its subspace consisting of elements

[4] Strictly speaking in order to be PSD matrix of a regular stationary process a matrix function on $[0, 2\pi]$ must satisfy other mild technical conditions (see [62] for details).

that are full-rank for almost every $\omega \in [0, 2\pi]$ by $\mathcal{P}_p^+$. Most of the literature prior to 2000 is devoted to geometrization of $\mathcal{P}_1^+$.

Remark 3 It is worth mentioning that the distances we discuss below here are blind to correlations, meaning that two processes might be correlated but their distance can be large or they can be uncorrelated but their distance can be zero. For us the starting point is the identification of a zero-mean (Gaussian) process with its probability distribution and hence its PSD. Consider the 1D case for convenience. Then in the Hilbert space geometry a distance between processes $\boldsymbol{y}_t^1$ and $\boldsymbol{y}_t^2$ can be defined as $\mathbb{E}\{(\boldsymbol{y}_t^1 - \boldsymbol{y}_t^2)^2\}$ in which case the correlation appears in the distance and a zero distance means almost surely equal sample paths, whereas in PSD-induced distances $\boldsymbol{y}_t$ and $-\boldsymbol{y}_t$ which have completely different sample paths have zero distance. In a more technical language, the topology induced by the PSD-induced distances on stochastic processes is coarser than the Hilbert space topology. Hence, perhaps to be more accurate we should further qualify the distances in this paper by the qualifier "PSD-induced". Obviously, the Hilbert space topology may be too restrictive in some practical applications. Interestingly, in the derivation of the Hellinger distance (see below) based on the optimal transport principle the issue of correlation shows up and there the optimality is achieved when the two processes are uncorrelated (hence the distance is computed as if the processes are uncorrelated, see [27, p. 292] for details). In fact, this idea is also present in our approach (and most of the other approaches), where in order to compare two LDSs we assume that they are stimulated with *the same* input process, meaning uncorrelated input processes with identical probability distributions (see Sect. 8.3).

The space $\mathcal{P}_p$ is an *infinite dimensional* cone which also has a convex *linear* structure coming from matrix addition and multiplication by nonnegative reals. The most immediate distance on this space is the standard Euclidean distance:

$$d_{\mathrm{E}}^2(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \|P_{\boldsymbol{y}^1}(\omega) - P_{\boldsymbol{y}^2}(\omega)\|^2 \mathrm{d}\omega, \tag{8.2}$$

where $\|\cdot\|$ is a matrix norm (e.g., the Frobenius norm $\|\cdot\|_F$). In the 1-dimensional case (i.e., $\mathcal{P}_1$) one could also define a distance based on the principle of *optimal decoupling* or *optimal (mass) transport* between the probability distributions of the two processes [27, p. 292]. This approach results in the formula:

$$d_{\mathrm{H}}^2(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \left|\sqrt{P_{\boldsymbol{y}^1}(\omega)} - \sqrt{P_{\boldsymbol{y}^2}(\omega)}\right|^2 \mathrm{d}\omega \tag{8.3}$$

This distance is derived in [28] and is also called the $\bar{d}_2$-distance (see also [27, p. 292]). In view of the Hellinger distance between probability measures [9], the above distance, in the literature, is also called the Hellinger distance [23]. Interestingly, d_{H} remains valid as the optimal transport-based distance for certain non-Gaussian processes, as well [27, p. 292]. The extension of the optimal transport-based definition to higher dimensions is not straightforward. However, note that in $\mathcal{P}_1$, d_{H} can be

318735_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:27/3/2014 Pages: 24 Layout: T1-Standard

thought of as a square root version of d_{E}. In fact, the square root based definition can be easily extended to higher dimensions, e.g., in (8.3) one could simply replace the scaler square roots with the (matrix) Hermitian square roots of $P_{\boldsymbol{y}^i}(\omega)$, $i = 1, 2$ (at each frequency ω) and use a matrix norm. Recall that the Hermitian square root of the Hermitian matrix Y is the unique Hermitian solution of the equation $Y = XX^H$, where H denotes conjugate transpose. We denote the Hermitian square root of Y as $Y^{1/2}$. Therefore, we could define the Hellinger distance in higher dimensions as

$$d_{\mathrm{H}}^2(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \|P_{\boldsymbol{y}^1}^{1/2}(\omega) - P_{\boldsymbol{y}^2}^{1/2}(\omega)\|_F^2 \mathrm{d}\omega \tag{8.4}$$

However note that, for any unitary matrix U, $X = Y^{1/2}U$ is also a solution to $Y = XX^H$ (but not Hermitian if U differs from the intensity). This suggests that, one may be able to do better by finding the best unitary matrix $U(\omega)$ to minimize $\|P_{\boldsymbol{y}^1}^{1/2}(\omega) - P_{\boldsymbol{y}^2}^{1/2}(\omega)U(\omega)\|_F$ (at each frequency ω). In [23] this idea has been used to define the (improved) Hellinger distance on $\mathcal{P}_p$ which can be written in closed-form as

$$d_{\mathrm{H}'}^2(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \|P_{\boldsymbol{y}^1}^{1/2} - P_{\boldsymbol{y}^2}^{1/2}\big(P_{\boldsymbol{y}^2}^{1/2}P_{\boldsymbol{y}^1}P_{\boldsymbol{y}^2}^{1/2}\big)^{-1/2}P_{\boldsymbol{y}^2}^{1/2}P_{\boldsymbol{y}^1}^{1/2}\|_F^2 \mathrm{d}\omega, \tag{8.5}$$

where dependence of the terms on ω has been dropped. Notice that the matrix $U(\omega) = \big(P_{\boldsymbol{y}^2}^{1/2}P_{\boldsymbol{y}^1}P_{\boldsymbol{y}^2}^{1/2}\big)^{-1/2}P_{\boldsymbol{y}^2}^{1/2}P_{\boldsymbol{y}^1}^{1/2}$ is unitary for every ω and in fact it is a transfer function of an all-pass possibly infinite dimensional linear filter [23]. Here, by an *all-pass* transfer function or filter $U(\omega)$ we mean exactly one for which $U(\omega)U(\omega)^H = I_p$. Also note that (8.5) seemingly breaks down if either of the PSDs is not full-rank; however, in fact, solving the related optimization shows that by continuity the expression remains valid. We should point out that recently a class of distances on $\mathcal{P}_1$ has been introduced by Georgiou et. al. based on the notion of optimal mass transport or morphism between PSDs (rather than probability distributions, as above) [25]. Such distances enjoy some nice properties, e.g., in terms of robustness with respect to multiplicative and additive noise [25]. An extension to $\mathcal{P}_p$ also has been proposed [53]; however, the extension is no longer a distance and it is not clear if it inherits the robustness property.

Another (possibly deeper) aspect of working with the square root of the PSD is about the ideas of *spectral factorization* and the *innovations process*. We review some basics which can be found e.g., in [6, 31, 32, 38, 62, 65]. The important fact is that the PSD $P_{\boldsymbol{y}}(\omega)$ of a regular process $\boldsymbol{y}_t$ in $\mathcal{P}_p$ is of constant rank $m \leq p$ almost everywhere in $[0, 2\pi]$. Moreover, it admits a factorization of the form $P_{\boldsymbol{y}}(\omega) = P_{l\boldsymbol{y}}(\omega)P_{l\boldsymbol{y}}(\omega)^H$, where $P_{l\boldsymbol{y}}(\omega)$ is $p \times m$-dimensional and uniquely determines its *analytic* extension $P_{l\boldsymbol{y}}(z)$ outside the unit disk in $\mathbb{C}$. In this factorization, $P_{l\boldsymbol{y}}(\omega)$, itself, is not determined uniquely and any two such factors are related by an $m \times m$-dimensional all-pass filter. However, if we require the extension $P_{l\boldsymbol{y}}(z)$ to be in the class of *minimum phase* filters, then the choice of the factor $P_{l\boldsymbol{y}}(\omega)$ becomes unique up to a constant unitary matrix.

318735_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:27/3/2014 Pages: 24 Layout: T1-Standard

A $p \times m$ $(m \leq p)$ transfer function matrix $H(z)$ is called minimum phase if it is analytic outside the unit disk and of constant rank m there (including at $z = \infty$). Such a filter has an inverse filter which is asymptotically stable. We denote this particular factor of $P_{\boldsymbol{y}}$ by $P_{+\boldsymbol{y}}$ and call it the *canonical spectral* factor. The canonical factor is still not unique, but the ambiguity is only in a constant $m \times m$ unitary matrix. The consequence is that $\boldsymbol{y}_t$ can be written as $\boldsymbol{y}_t = \sum_{\tau=0}^{\infty} \mathbf{p}_{+\tau} \boldsymbol{\epsilon}_{t-\tau}$, where the $p \times m$ matrix sequence $\{\mathbf{p}_{+t}\}_{t=0}^{\infty}$ is the inverse Fourier transform of $P_{+\boldsymbol{y}}(\omega)$ and $\boldsymbol{\epsilon}_t$ is an m-dimensional *white* noise process with covariance I_m. This means that $\boldsymbol{y}_t$ is the output of a linear filter (i.e., an LDS of possibly *infinite* order) excited by a white noise process with standard covariance. The process $\boldsymbol{\epsilon}_t$ is called the *innovations process* or *fundamental process* of $\boldsymbol{y}_t$. Under Gaussian assumption the innovations process is determined uniquely, otherwise it is determined up to an $m \times m$ unitary factor. The important case is when $P_{\boldsymbol{y}}(z)$ is full-rank outside the unit disk, in which case the inverse filter $P_{+\boldsymbol{y}}^{-1}$ is well-defined and asymptotically stable, and one could recover the innovations process by filtering $\boldsymbol{y}_t$ by its *whitening* filter $P_{+\boldsymbol{y}}^{-1}$.

Now, to compare two processes one could, obviously, somehow compare their canonical spectral factors[5] or if they are in $\mathcal{P}_p^+$ their whitening filters. In [38] a large class of divergences based on the idea of comparing associated whitening filters (in the frequency domain) have been proposed. For example, let $P_{+\boldsymbol{y}_i}$ be the canonical factor of $P_{\boldsymbol{y}_i}$, $i = 1, 2$. If one filters $\boldsymbol{y}_t^i$, $i = 1, 2$, with $P_{+\boldsymbol{y}^j}^{-1}$, $j = 1, 2$, then the output PSD is $P_{+\boldsymbol{y}^j}^{-1} P_{\boldsymbol{y}^i} P_{+\boldsymbol{y}^j}^{-H}$. Note that when $i = j$ then the output PSD is I_p the $p \times p$ identity across every frequency. It can be shown that $d_I(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \mathrm{tr}(P_{+\boldsymbol{y}^1}^{-1} P_{\boldsymbol{y}^2} P_{+\boldsymbol{y}^1}^{-H} - I_p) + \mathrm{tr}(P_{+\boldsymbol{y}^2}^{-1} P_{\boldsymbol{y}^1} P_{+\boldsymbol{y}^2}^{-H} - I_p) \mathrm{d}\omega$ is a symmetric divergence [38]. Note that $d_I(\boldsymbol{y}^1, \boldsymbol{y}^2)$ is independent of the unitary ambiguity in the canonical factor and in fact

$$d_I(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \mathrm{tr}(P_{\boldsymbol{y}^1}^{-1} P_{\boldsymbol{y}^2} + P_{\boldsymbol{y}^2}^{-1} P_{\boldsymbol{y}^1} - 2I_p) \mathrm{d}\omega. \tag{8.6}$$

Such divergences enjoy certain invariance properties e.g., if we filter both processes with a common minimum phase filter, then the divergence remains unchanged. In particular, it is scale-invariant. Such properties are shared by the distances or divergences that are based on the ratios of PSDs (see below for more examples). Scale invariance in the case of 1D PSDs has been advocated as a desirable property, since in many cases the shape of the PSDs rather than their relative scale is the discriminative feature (see e.g., [9, 26]).

One can arrive at similar distances from other geometric or probabilistic paths. One example is the famous Itakura-Saito divergence (sometimes called distance) between PSDs in $\mathcal{P}_1^+$ which is defined as

[5] In fact, our approach (in Sects. 8.3–8.5) is also based on the idea of comparing the minimum phase (i.e., canonical) filters or factors in the case of processes with rational spectra. However, instead of comparing the associated transfer functions or impulse responses we try to compare the associated state-space realizations (in a specific sense). This approach, therefore, is in some sense *structural* or *generative*, since it tries to compare how the processes are generated (according to the state-space representation) and the model order plays an explicit role in it.

$$d_{\mathrm{IS}}(\boldsymbol{y}^1,\boldsymbol{y}^2)=\int\left(\frac{P_{\boldsymbol{y}^1}}{P_{\boldsymbol{y}^2}}-\log\frac{P_{\boldsymbol{y}^1}}{P_{\boldsymbol{y}^2}}-1\right)\mathrm{d}\omega \tag{8.7}$$

This divergence has been used in practice, at least, since the 1970s (see [48] for references). The Itakura-Saito divergence can be derived from the Kullback-Leibler divergence between (infinite dimensional) probability densities of the two processes (The definition is a time-domain based definition, however, the final result is readily expressible in the frequency domain).[6] On the other hand, Amari's information geometry-based approach [5, Chap. 5] allows to geometrize $\mathcal{P}_1^+$ in various ways and yields different distances including the Itakura-Saito distance (8.7) or a Riemannian distance such as

$$d_{\mathrm{R}}^2(\boldsymbol{y}^1,\boldsymbol{y}^2)=\int\left(\log\left(\frac{P_{\boldsymbol{y}^1}}{P_{\boldsymbol{y}^2}}\right)\right)^2\mathrm{d}\omega. \tag{8.8}$$

Furthermore, in this framework one can define geodesics between two processes under various Riemannian or non-Riemannian *connections*. The high-dimensional version of the Itakura-Saito distance has also been known since the 1980s [42] but less used in practice:

$$d_{\mathrm{IS}}(\boldsymbol{y}^1,\boldsymbol{y}^2)=\int\left(\mathrm{trace}(P_{\boldsymbol{y}^2}^{-1}P_{\boldsymbol{y}^1})-\log\det P_{\boldsymbol{y}^2}^{-1}P_{\boldsymbol{y}^1}-p\right)\mathrm{d}\omega \tag{8.9}$$

Recently, in [38] a Riemannian framework for geometrization of $\mathcal{P}_p^+$ for $p\geq 1$ has been proposed, which yields Riemannian distances such as:

$$d_{\mathrm{R}}^2(\boldsymbol{y}^1,\boldsymbol{y}^2)=\int\|\log\left(P_{\boldsymbol{y}^1}^{-1/2}P_{\boldsymbol{y}^2}P_{\boldsymbol{y}^1}^{-1/2}\right)\|_F^2\mathrm{d}\omega, \tag{8.10}$$

where log is the standard matrix logarithm. In general, such approaches are not suited for large p due to computational costs and the full-rankness requirement. We should stress that in (very) high dimensions the assumption of full-rankness of PSDs is not a viable one, in particular because usually not only the actual time series are highly correlated but also the contaminating noises are correlated, as well. In fact, this has lead to the search for models capturing this quality. One example is the class of *generalized linear dynamic factor models*, which are closely related to the tall, full rank LDS models (see Sect. 8.3.3 and [20, 24]).

The above mentioned issues aside, for the purposes of Problem 1, the space $\mathcal{P}_p$ (or even $\mathcal{P}_p^+$) is *too large*. The reason is that it includes, e.g., ARMA processes of arbitrary large orders, and it is not clear, e.g., how an *average* of some ARMA models

[6] Notice that defining distances between probability densities in the time domain is a more general approach than the PSD-based approaches, and it can be employed in the case of nonstationary as well as non-Gaussian processes. However, such an approach, in general, is computationally difficult.

or processes of equal order might turn out. As mentioned before, it is convenient or reasonable to require the average to be of the same order.[7]

8.2.2 Geometrizing the Spaces of Models

Any distance on $\mathcal{P}_p$ (or $\mathcal{P}_p^+$) induces a distance, e.g., on a subspace corresponding to AR or ARMA models of a fixed order. This is an example of an extrinsic distance induced from an *infinite dimensional ambient* space to a *finite dimensional subspace*. In general, this framework is not ideal and we might try to, e.g., define an *intrinsic* distance on the finite dimensional subspace. In fact, Amari's original paper [4] lays down a framework for this approach, but lacks actual computations. For the one-dimensional case in [61], based on Amari's approach, distances between models in the space of ARMA models of fixed order are derived. For high order models or in high dimensions, such calculations are, in general, computationally difficult [61]. The main reason is that the dependence of PSD-based distances on state-space or ARMA parameters, in general, is highly nonlinear (the important exception is for parameters of AR models, especially in 1D).

Alternative approaches also have been pursued. For example, in [57] the main idea is to compare (based on the ℓ^2 norm) the coefficients of the infinite order AR models of two processes. This is essentially the same as comparing (in time domain) the whitening filters of the two processes. This approach is limited to $\mathcal{P}_p^+$ and computationally demanding for large p. See [19] for examples of classification and clustering of 1D time-series using this approach. In [8], the space of 1D AR processes of a fixed order is geometrized using the geometry of positive-definite Toeplitz matrices (via the reflection coefficients parameterization), and, moreover, L^p averaging on that space is studied. In [50] a (pseudo)-distance between two processes is defined through a weighted ℓ^2 distance between the (infinite) sequences of the *cepstrum* coefficients of the two processes. Recall that the cepstrum of a 1D signal is the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of the signal. In the frequency domain this distance (known as the Martin distance) can be written as (up to a multiplicative constant)

$$d_{\mathrm{M}}^2(\boldsymbol{y}_1, \boldsymbol{y}_2) = \int \left(\mathfrak{D}^{\frac{1}{2}} \log \left(\frac{P_{\boldsymbol{y}_1}}{P_{\boldsymbol{y}_2}} \right) \right)^2 \mathrm{d}\omega, \tag{8.11}$$

where $\mathfrak{D}^{\lambda}$ is the fractional derivative operator in the frequency domain interpreted as multiplication of the corresponding Fourier coefficients in the time domain by $e^{\pi i \lambda/2} n^{\lambda}$ for $n \geq 0$ and by $e^{-\pi i \lambda/2}(-n)^{\lambda}$ for $n < 0$. Notice that d_M is scale-invariant in the sense described earlier and also it is a pseudo-distance since it is zero if the PSDs are multiple of each other (this is a true scale-invariance property, which in

[7] Interestingly, for an average defined based on the Itakura-Saito divergence in the space of 1D AR models this property holds [26], see also [5, Sect. 5.3].

certain applications is highly desirable).[8] Interestingly, in the case of 1D ARMA models, d_M can be expressed conveniently in closed form in terms of the poles and zeros of the models [50]. Moreover, in [18] it is shown that d_M can be calculated quite efficiently in terms of the parameters of the state-space representation of the ARMA processes. In fact, the Martin distance has a simple interpretation in terms of the subspace angles between the extended observability matrices (cf. Sect. 8.4.3) of the state-space representations [18]. This brings about important computational advantages and has allowed to extend a form of Martin distance to higher dimensions (see e.g., [16]). However, it should be noted that the extension of the Martin distance to higher dimensions in such a way that all its desirable properties carry over has proven to be difficult [13].[9] Nevertheless, some extensions have been quite effective in certain high-dimensional applications e.g., video classification [16]. In [16], the approach of [18] is shown to be a special case of the family of Binet-Cauchy kernels introduced in [64], and this might explain the effectiveness of the extensions of the Martin distance to higher dimensions.

In summary, we should say that the extensions of the geometrical methods discussed in this section to $\mathcal{P}_p$ for $p > 1$ do not seem obvious or otherwise they are computationally very expensive. Moreover, these approaches often yield extrinsic distances induced from infinite dimensional ambient spaces, which e.g., in the case of averaging LDSs of *fixed* order can be problematic.

8.2.3 Control-Theoretic Approaches

More relevant to us are [33, 46], where (*intrinsic*) state-space based Riemannian distances between LDSs of fixed size and fixed order have been studied. Such approaches ideally suit Problem 1, but they are computationally demanding. More recently, in [1] and subsequently in [2, 3] we introduced group action induced distances on certain spaces of LDSs of *fixed size* and *order*. As it becomes clear in the following, an important feature of this approach is that the LDS order is *explicit* in the construction of the distance, and the state-space parameters appear in the distance in a simple form. These features make certain related calculations (e.g., optimization) much more convenient (compared with other methods). Another aspect of our approach is that, contrary to most of the distances discussed so far which compare the PSDs or

[8] It is interesting to note that by a simple modification some of the spectral-ratio based distances can attain this property, e.g., by modifying d_{R} in (8.8) as $d_{\mathrm{RI}}^2(\boldsymbol{y}^1, \boldsymbol{y}^2) = \int \big(\log\big(\frac{P_{y^1}}{P_{y^2}}\big)\big)^2 \mathrm{d}\omega - \big(\int \log\big(\frac{P_{y^1}}{P_{y^2}}\big)\mathrm{d}\omega\big)^2$ (see also [9, 25, 49]).

[9] This and the results in [53] underline the fact that defining distances on $\mathcal{P}_p$ for $p > 1$ may be challenging, not only from a computational point of view but also from a theoretical one. In particular, certain nice properties in 1D do not automatically carry over to higher dimensions by simple extension of definitions in 1D.

the canonical factors directly, our approach amounts to comparing the generative or the structural models of the processes or how they are generated. This feature also could be useful in designing more application-specific or structure-aware distances.

8.3 Processes Generated by LDSs of Fixed Order

Consider an LDS, M, of the form (8.1) with *a* realization $R = (A, B, C, D) \in \widetilde{\mathcal{SL}}_{m,n,p}$.[10] In the sequel, for various reasons, we will restrict ourselves to increasingly smaller submanifolds of $\widetilde{\mathcal{SL}}_{m,n,p}$ which will be denoted by additional superscripts. Recall that the $p \times m$ matrix transfer function is $T(z) = D + C(I_n - z^{-1}A)^{-1}B$, where $z \in \mathbb{C}$ and I_n is the n-dimensional identity matrix. *We assume that all LDSs are excited by the standard white Gaussian process.* Hence, the output PSD matrix (in the z-domain) is the $p \times p$ matrix function $P(z) = T(z)T^\top(z^{-1})$. The PSD is a rational matrix function of z whose rank (a.k.a. *normal rank*) is constant almost everywhere in $\mathbb{C}$. Stationarity of the output process is guaranteed if M is asymptotically stable. We denote the submanifold of such realizations by $\widetilde{\mathcal{SL}}^{\mathrm{a}}_{m,n,p} \subset \widetilde{\mathcal{SL}}_{m,n,p}$.

8.3.1 Embedding Stochastic Processes in LDS Spaces

Two (stochastic) LDSs are indistinguishable if their output PSDs are equal. Using this equivalence on the entire set of LDSs is not useful, because, as mentioned earlier two transfer functions which differ by an all-pass filter result in the same PSD. Therefore, the equivalence relation could induce a complicated many-to-one correspondence between the LDSs and the subspace of stochastic processes they generate. However, if we restrict ourselves to the subspace of minimum phase LDSs the situation improves. Let us denote the subspace of minimum-phase realizations by $\widetilde{\mathcal{SL}}^{\mathrm{a,mp}}_{m,n,p} \subset \widetilde{\mathcal{SL}}^{\mathrm{a}}_{m,n,p}$. This is clearly an open submanifold of $\widetilde{\mathcal{SL}}^{\mathrm{a}}_{m,n,p}$. In $\widetilde{\mathcal{SL}}^{\mathrm{a,mp}}_{m,n,p}$, the canonical spectral factorization of the output PSD is unique up to an orthogonal matrix [6, 62, 65]: let $T_1(z)$ and $T_2(z)$ have realizations in $\widetilde{\mathcal{SL}}^{\mathrm{a,mp}}_{m,n,p}$ and let $T_1(z)T_1^\top(z^{-1}) = T_2(z)T_2^\top(z^{-1})$, then $T_1(z) = T_2(z)\Theta$ for a unique $\Theta \in O(m)$, where $O(m)$ is the Lie group of $m \times m$ orthogonal matrices. Therefore, any p-dimensional processes with PSD of normal rank m can be identified with a simple equivalent class of stable and minimum-phase transfer functions and the corresponding LDSs.[11]

[10] It is crucial to have in mind that we explicitly distinguish between the LDS, M, and its realization R, which is not unique. As it becomes clear soon, an LDS has an equivalent class of realizations.

[11] These rank conditions, interestingly, have differential geometric significance in yielding nice quotient spaces, see Sect. 8.4.

8.3.2 Equivalent Realizations Under Internal and External Symmetries

A fundamental fact is that there are *symmetries* or *invariances* due to certain Lie *group actions* in the model (8.1). Let $GL(n)$ denote the Lie group of $n \times n$ non-singular (real) matrices. We say that the Lie group $GL(n) \times O(m)$ *acts* on the realization space $\widetilde{\mathcal{SL}}_{m,n,p}$ (or its subspaces) via the action $\bullet$ defined as[12]

$$(P, \Theta) \bullet (A, B, C, D) = (P^{-1}AP, P^{-1}B\Theta, CP, D\Theta). \tag{8.12}$$

One can easily verify that under this action the output covariance sequence (or PSD) remains invariant. In general, the *converse* is not true. That is, two output covariance sequences might be equal while their corresponding realizations are not related via $\bullet$ (due to non-minimum phase and the action not being *free* [47], also see below). Recall that the action of a group on a set is called free if every element of the set is only fixed by the identity element of the group. For the converse to hold we need to impose further *rank* conditions, as we see next.

8.3.3 From Processes to Realizations (The Rank Conditions)

Now, we study some rank conditions (i.e., submanifolds of $\widetilde{\mathcal{SL}}_{m,n,p}$ on) which $\bullet$ is a free action.

8.3.3.1 Observable, Controllable, and Minimal Realizations

Recall that *controllability* and *observability* matrices of order k are defined as $\mathcal{C}_k = [B, AB, \ldots, A^{k-1}B]$ and $\mathcal{O}_k = [C^\top, (CA)^\top, \ldots, (CA^{k-1})^\top]^\top$, respectively. A realization is called *controllable* (resp. *observable*) if $\mathcal{C}_k$ (resp. $\mathcal{O}_k$) is of rank n for $k = n$. We denote the subspace of controllable (resp. observable) realizations by $\widetilde{\mathcal{SL}}^{\mathrm{co}}_{m,n,p}$ (resp. $\widetilde{\mathcal{SL}}^{\mathrm{ob}}_{m,n,p}$). The space $\widetilde{\mathcal{SL}}^{\min}_{m,n,p} = \widetilde{\mathcal{SL}}^{\mathrm{co}}_{m,n,p} \cap \widetilde{\mathcal{SL}}^{\mathrm{ob}}_{m,n,p}$ is called the space of *minimal* realizations. An important fact is that we cannot reduce the order (i.e., the size of A) of a minimal realization without changing its input-output behavior.

8.3.3.2 Tall, Full Rank LDSs

Another (less studied) rank condition is when C is of rank n (here $p \geq n$ is required). Denote by $\widetilde{\mathcal{SL}}^{\mathrm{tC}}_{m,n,p} \subset \widetilde{\mathcal{SL}}^{\mathrm{ob}}_{m,n,p}$ the subspace of such realizations and call a corresponding LDS *tall and full-rank*. Such LDSs are closely related to generalized linear

[12] Strictly speaking $\bullet$ is a right action; however, it is notationally convenient to write it as a left action in (8.12).

dynamic factor models for (very) high-dimensional time series [20] and also appear in video sequence modeling [1, 12, 60]. It is easy to verify that all the above realization spaces are smooth open submanifolds of $\widetilde{\mathcal{SL}}_{m,n,p}$. Their corresponding submanifolds of stable or minimum-phase LDSs (e.g., $\widetilde{\mathcal{SL}}^{\mathrm{a,mp,co}}_{m,n,p}$) are defined in an obvious way.

The following proposition forms the basis of our approach to defining distances between processes: any distance on the space of LDSs with realizations in the above submanifolds (with rank condition) can be used to define a distance on the space of processes generated by those LDSs.

Proposition 1 *Let $\tilde{\Sigma}_{m,n,p}$ be $\widetilde{\mathcal{SL}}^{a,mp,co}_{m,n,p}$, $\widetilde{\mathcal{SL}}^{a,mp,ob}_{m,n,p}$, $\widetilde{\mathcal{SL}}^{a,mp,min}_{m,n,p}$, or $\widetilde{\mathcal{SL}}^{a,mp,tC}_{m,n,p}$. Consider two realizations $R_1, R_2 \in \tilde{\Sigma}_{m,n,p}$ excited by the standard white Gaussian process. Then we have:*

1. *If $(P,\Theta)\bullet R_1 = R_2$ for some $(P,\Theta) \in GL(n)\times O(m)$, then the two realizations generate the same (stationary) generate output process (i.e., outputs have the same PSD matrices).*
2. *Conversely, if the outputs of the two realizations are equal (i.e., have the same PSD), then there exists a unique $(P,\Theta) \in GL(n)\times O(m)$ such that $(P,\Theta)\bullet R_1 = R_2$.*

8.4 Principal Fiber Bundle Structures Over Spaces of LDSs

As explained above, an LDS, M, has an equivalent class of realizations related by the action $\bullet$. Hence, M sits naturally in a *quotient* space, namely $\widetilde{\mathcal{SL}}_{m,n,p}/(GL(n)\times O(m))$. However, this quotient space is not smooth or even Hausdorff. Recall that if a Lie group G acts on a manifold *smoothly*, *properly*, and *freely*, then the quotient space has the structure of a *smooth manifold* [47]. Smoothness of $\bullet$ is obvious. In general, the action of a *non-compact* group such as $GL(n)\times O(m)$ is *not* proper. However, one can verify that the rank conditions we imposed in Proposition 1 are enough to make $\bullet$ both a proper and free action on the realization submanifolds (see [2] for a proof). The resulting quotient manifolds are denoted by dropping the superscript $\sim$, e.g., $\mathcal{SL}^{\mathrm{a,mp,min}}_{m,n,p}$. The next theorem, which is an extension of existing results, e.g., in [33] shows that, in fact, we have a principal fiber bundle structure.

Theorem 1 *Let $\tilde{\Sigma}_{m,n,p}$ be as in Proposition 1 and $\Sigma_{m,n,p} = \tilde{\Sigma}_{m,n,p}/(GL(n)\times O(m))$ be the corresponding quotient LDS space. The realization-system pair $(\tilde{\Sigma}_{m,n,p}, \Sigma_{m,n,p})$ has the structure of a smooth principal fiber bundle with structure group $GL(n)\times O(m)$. In the case of $\mathcal{SL}^{a,mp,tC}_{m,n,p}$ the bundle is trivial (i.e., diffeomorphic to a product), otherwise it is trivial only when $m = 1$ or $n = 1$.*

The last part of the theorem has an important consequence. Recall that a principal bundle is trivial if it diffeomorphic to global product of its base space and its structure group. Equivalently, this means that a trivial bundle admits a global smooth cross section or what is known as a smooth canonical form in the case of LDSs, i.e., a globally smooth mapping $s\colon \Sigma_{m,n,p} \rightarrow \widetilde{\Sigma}_{m,n,p}$. This theorem implies that the minimality condition is a complicated nonlinear constraint, in the sense that it makes the bundle twisted and nontrivial for which no continuous canonical form exists. Establishing this obstruction put an end to control theorists' search for canonical forms for MIMO LDSs in the 1970s and explained why system identification for MIMO LDSs is a challenging task [11, 15, 36].

On the other hand, one can verify that $(\widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p}, \mathcal{SL}^{\text{a,mp,tC}}_{m,n,p})$ is a trivial bundle. Therefore, for such systems global canonical forms exist and they can be used to define distances, i.e., if $s : \mathcal{SL}^{\text{a,mp,tC}}_{m,n,p} \rightarrow \widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p}$ is such a canonical form then $d_{\mathcal{SL}^{\text{a,mp,tC}}_{m,n,p}}(M_1, M_2) = \tilde{d}_{\widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p}}(s(M_1), s(M_2))$ defines a distance on $\mathcal{SL}^{\text{a,mp,tC}}_{m,n,p}$ for any distance $\tilde{d}_{\widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p}}$ on the realization space. In general, unless one has some specific knowledge there is no preferred choice for section or canonical form. If one has a *group-invariant* distance on the realization space, then the distance induced from using a cross section might be inferior to the *group action induced distance*, in the sense it may result in artificially larger distance. In the next section we review the basic idea behind group action induced distances in our application.

8.4.1 Group Action Induced Distances

Figure 8.1a schematically shows a realization bundle $\widetilde{\Sigma}$ and its base LDS space Σ. Systems $M_1, M_2 \in \Sigma$ have realizations R_1 and R_2 in $\widetilde{\Sigma}$, respectively. Let us assume that a $G = GL(n) \times O(n)$-invariant distance $\tilde{d}_G$ on the realization bundle is given. The realizations, R_1 and R_2, in general, are not aligned with each other, i.e., $\tilde{d}_G(R_1, R_2)$ can be still reduced by sliding one realization along its fiber as depicted in Fig. 8.1b. This leads to the definition of the group action induced distance:[13]

$$d_{\Sigma}(M_1, M_2) = \inf_{(P,\Theta)\in G} \tilde{d}_{\tilde{\Sigma}}((P, \Theta) \bullet R_1, R_2) \tag{8.13}$$

In fact, one can show that $d_{\Sigma}(\cdot, \cdot)$ is a true distance on Σ, i.e., it is symmetric and positive definite and obeys the triangle inequality (see e.g., [66]).[14]

The main challenge in the above approach is the fact that, due to non-compactness of $GL(n)$, constructing a $GL(n) \times O(n)$-invariant distance is computationally difficult. The construction of such a distance can essentially be accomplished by defining a $GL(n) \times O(n)$-invariant Riemannian on the realization space and solving

[13] We may call this an alignment distance. However, based on the same principle in Sect. 8.5 we define another group action induced distance, which we explicitly call the alignment distance. Since our main object of interest is that distance, we prefer not to call the distance in (8.13) an alignment distance.

[14] It is interesting to note that some of the good properties of the k-nearest neighborhood algorithms on a general metric space depend on the triangle inequality [21].

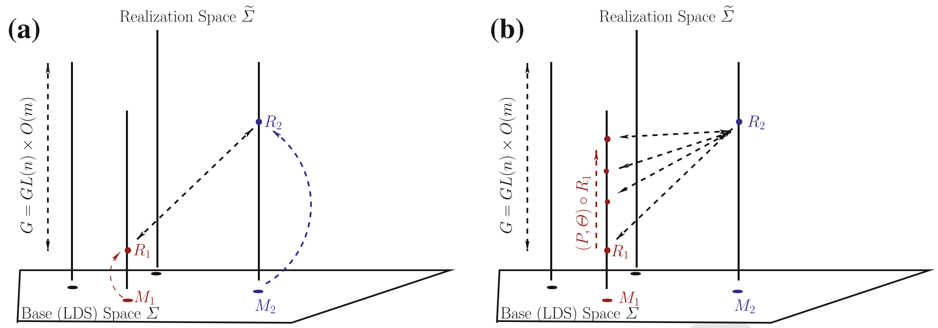

Fig. 8.1 Over each LDS in Σ sits a realization fiber. The fibers together form the realization space (bundle) $\widetilde{\Sigma}$. If given a G-invariant distance on the realization bundle, then one can define a distance on the LDS space by aligning any realizations R_1, R_2 of the two LDSs M_1, M_2 as in (8.13)

the corresponding geodesic equation, as well as searching for global minimizers.[15] Such a Riemannian metric for deterministic LDSs was proposed in [45, 46]. One could also start from (an already invariant) distance on a large ambient space such as $\mathcal{P}_p$ and specialize it to the desired submanifold Σ of LDSs to get a Riemannian manifold on Σ and then thereon solve geodesic equations, etc. to get an *intrinsic* distance (e.g., as reported in [33, 34]). Both of these approaches seem very complicated to implement for the case of very high-dimensional LDSs. Instead, our approach is to use *extrinsic* group action induced distances, which are induced from unitary-invariant distances on the realization space. For that we recall the notion of reduction of structure group on a principal fiber bundle.

8.4.2 Standardization: Reduction of the Structure Group

Next, we recall the notion of reducing a bundle with non-compact structure group to one with a compact structure group. This will be useful in our geometrization approach in the next section. Interestingly, bundle reduction also appears in *statistical analysis of shapes* under the name of *standardization* [43]. The basic fact is that any principal fiber G-bundle $(\tilde{\Sigma}, \Sigma)$ can be *reduced* to an OG-subbundle $\widetilde{\mathcal{O}\Sigma} \subset \tilde{\Sigma}$, where OG is the maximal compact subgroup of G [44]. This reduction means that Σ is *diffeomorphic* to $\widetilde{\mathcal{O}\Sigma}/OG$ (i.e., no *topological* information is lost by going to the subbundle and the subgroup). Therefore, in our cases of interest we can reduce a $GL(n) \times O(m)$-bundle to an $OG(n, m) = O(n) \times O(m)$-subbundle. We call such a subbundle a *standardized* realization space or (sub)bundle. One can perform reduction to various standardized subbundles and there is *no* canonical reduction.

[15] This problem, in general, is difficult, among other things, because it is a non-convex (infinite-dimensional) variational problem. Recall that in Riemannian geometry the non-convexity of the arc length variational problem can be related to the non-trivial topology of the manifold (see e.g., [17]).

318735_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:27/3/2014 Pages: 24 Layout: T1-Standard

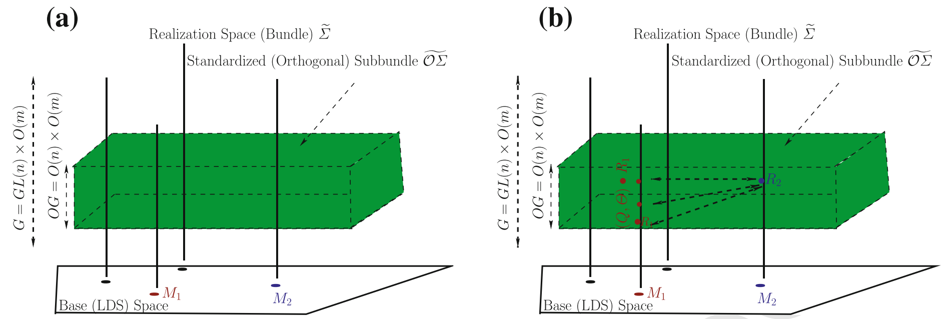

Fig. 8.2 A standardized subbundle $\widetilde{\mathcal{O}\Sigma}_{m,n,p}$ of $\widetilde{\Sigma}_{m,n,p}$ is a subbundle on which G acts via its *compact* subgroup OG. The quotient space $\widetilde{\mathcal{O}\Sigma}_{m,n,p}/OG$ still is diffeomorphic to the base space $\widetilde{\Sigma}_{m,n,p}$. One can define an alignment distance on the base space by aligning realizations $R_1, R_2 \in \widetilde{\mathcal{O}\Sigma}_{m,n,p}$ of $M_1, M_2 \in \Sigma_{m,n,p}$ as (8.15)

However, in each application one can choose an *interesting* one. A reduction is in spirit similar to the Gram-Schmidt orthonormalization [44, Chap. 1]. Figure 8.2a shows a standardized subbundle $\widetilde{\mathcal{O}\Sigma}$ in the realization bundle $\widetilde{\Sigma}$.

8.4.3 Examples of Realization Standardization

As an example consider $R = (A, B, C, D) \in \widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p}$, and let $C = UP$ be an orthonormalization of C, where $U^\top U = I_n$ and $P \in GL(n)$. Now the new realization $\hat{R} = (P^{-1}, I_m) \bullet R$ belongs to the $O(n)$-subbundle $\widetilde{\mathcal{OSL}}^{\text{a,mp,tC}}_{m,n,p} = \{R \in \widetilde{\mathcal{SL}}^{\text{a,mp,tC}}_{m,n,p} | C^\top C = I_n\}$.

Other forms of bundle reduction, e.g., in the case of the nontrivial bundle $\widetilde{\mathcal{SL}}^{\text{a,mp,min}}_{m,n,p}$ are possible. In particular, via a process known as *realization balancing* (see [2, 37]), we can construct a large family of standardized subbundles. For example, a more sophisticated one is in the case of $\widetilde{\mathcal{SL}}^{\text{a,mp,min}}_{m,n,p}$ via the notion of (internal) *balancing*. Consider the symmetric $n \times n$ matrices $W_c = \mathcal{C}_\infty \mathcal{C}_\infty^\top$ and $W_o = \mathcal{O}_\infty^\top \mathcal{O}_\infty$, which are called controllability and observability Gramians, respectively, and where $\mathcal{C}_\infty$ and $\mathcal{O}_\infty$ are called extended controllability and observability matrices, respectively (see the definitions in Sect. 8.3.3.1 with $k = \infty$). Due to the minimality assumption, both W_o and W_c are positive definite. Notice that under the action $\bullet$, W_c transforms to $P^{-1} W_c P^{-\top}$ and W_o to $P^\top W_o P$. Consider the function $h: GL(n) \to \mathbb{R}$ defined as $h(P) = \text{trace}(P^{-1} W_c P^{-\top} + P^\top W_o P)$. It is easy to see that h is invariant on $O(n)$. More importantly, it can be shown that any critical point P_1 of h is global minimizer and if P_2 is any other minimizer then $P_1 = P_2 Q$ for some $Q \in O(n)$ [37]. Minimizing h is called balancing (in the sense of Helmke). One can show that balancing is, in fact, a standardization in the sense that we defined (a proof of this fact will appear elsewhere). Note that a more specific form of balancing called diagonal balancing (due to Moore [52]) is more common in the control literature, however, that cannot be considered as a form of reduction of the structure group. The interesting intuitive reason is that it tries to reduce the structure group beyond the orthogonal group to the identity element, i.e., to get a canonical form (see also [55]). However, it fails in the sense that, as mentioned above, it cannot give a *smooth* canonical form, i.e., a section which is diffeomorphic to $\mathcal{SL}^{\mathrm{a,mp,min}}_{m,n,p}$.

8.5 Extrinsic Quotient Geometry and the Alignment Distance

In this section, we propose to use the large class of *extrinsic* unitary invariant distances on *a* standardized realization subbundle to build distances on *the* LDS base space. The main benefits are that such distances are abundant, the ambient space is *not* too large (e.g., not infinite dimensional), and calculating the distance in the base space boils down to a static optimization problem (albeit non-convex). Specifically, let $\tilde{d}_{\widetilde{\mathcal{O}\Sigma}_{m,n,p}}$ be a unitary invariant distance on a standardized realization subbundle $\widetilde{\mathcal{O}\Sigma}_{m,n,p}$ with the base $\Sigma_{m,n,p}$ (as in Theorem 1). One example of such a distance is

$$\tilde{d}^2_{\widetilde{\mathcal{O}\Sigma}_{m,n,p}}(R_1,R_2)=\lambda_A\|A_1-A_2\|_F^2+\lambda_B\|B_1-B_2\|_F^2+\lambda_C\|C_1-C_2\|_F^2+\lambda_D\|D_1-D_2\|_F^2, \tag{8.14}$$

where $\lambda_A,\lambda_B,\lambda_C,\lambda_D>0$ are constants and $\|\cdot\|_F$ is the matrix Frobenius norm. A group action induced distance (called the *alignment* distance) between two LDSs $M_1, M_2 \in \Sigma_{m,n,p}$ with realizations $R_1, R_2 \in \widetilde{\mathcal{O}\Sigma}_{m,n,p}$ is found by solving the *realization alignment* problem (see Fig. 8.2b)

$$d^2_{\Sigma_{m,n,p}}(M_1,M_2)=\min_{(Q,\Theta)\in O(n)\times O(m)} \tilde{d}^2_{\widetilde{\mathcal{O}\Sigma}_{m,n,p}}\big((Q,\Theta)\bullet R_1, R_2\big). \tag{8.15}$$

In [39] a fast algorithm is developed which (with little modification) can be used to compute this distance.

Remark 4 We stress that, via the identification of a process with its canonical spectral factors (Proposition 1 and Theorem 1), $d_{\Sigma_{m,n,p}}(\cdot,\cdot)$ is (or induces) a distance on the space of processes generated by the LDSs in $\Sigma_{m,n,p}$. Therefore, in the sprit of distances studied in Sect. 8.2 we could have written $d_{\Sigma_{m,n,p}}(\boldsymbol{y}_1,\boldsymbol{y}_2)$ instead of $d_{\Sigma_{m,n,p}}(M_1,M_2)$, where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the processes generated by M_1 and M_2 when excited by the standard Gaussian process. However, the chosen notation seems more convenient.

Remark 5 Calling the static *global* minimization problem (8.15) "easy" in an absolute term is oversimplification. However, even this *global* minimization over orthogonal matrices is definitely simpler than solving the nonlinear geodesic ODEs and finding shortest geodesics globally (an infinite-dimensional dynamic programming problem). It is our ongoing research to develop fast and reliable algorithms to solve (8.15). Our experiments indicate that the Jacobi algorithm in [39] is quite effective in finding global minimizers.

In [1], this distance was first introduced on $\mathcal{SL}_{m,n,p}^{\mathrm{a,mp,tC}}$ with the standardized subbundle $\widetilde{\mathcal{OSL}}_{m,n,p}^{\mathrm{a,mp,tC}}$. The distance was used for efficient video sequence classification (using 1-nearest neighborhood and nearest mean methods) and clustering (e.g., via defining *averages* or a k-means like algorithm). However, it should be mentioned that in video applications (for reasons which are not completely understood) the comparison of LDSs based on the (A, C) part in (8.1) has proven quite effective (in fact, such distances are more commonly used than distances based on comparing the full model). Theretofore, in [1], the alignment distance (8.15) with parameters $\lambda_B = \lambda_D = 0$ was used, see (8.14). An algorithm called the *align and average* is developed to do averaging on $\mathcal{SL}_{m,n,p}^{\mathrm{a,mp,tC}}$ (see also [2]). One defines the average $\bar{M}$ of LDSs $\{M_i\}_{i=1}^N \subset \mathcal{SL}_{m,n,p}^{\mathrm{a,mp,tC}}$ (the so-called Fréchet mean or average) as a minimizer of the sum of the squares of distances:

$$\bar{M} = \mathrm{argmin}_M \sum_{i=1}^{N} d^2_{\mathcal{SL}_{m,n,p}^{\mathrm{a,mp,tC}}}(M, M_i). \tag{8.16}$$

The align and average algorithm is essentially and alternative minimization algorithm to find a solution. As a result, in each step it *aligns* the realizations of the LDSs M_i to that of the current estimated average, then a Euclidean *average* of the aligned realizations is found and afterwards the found C matrix is orthonormalized, and the algorithm iterates these steps till convergence (see [1, 2] for more details). A nice feature of this algorithms is that (generically) the average LDS $\bar{M}$ by construction will be of order n and minimum phase (and under certain conditions stable). An interesting question is whether the average model found this way is asymptotically stable, by construction. The answer most likely, in general, is negative. However, in a special case it can be positive. Let $\|A\|_2$ denote the 2-norm (i.e., the largest singular value) of the matrix A. In the case the standardized realizations $R_i \in \widetilde{\mathcal{OSL}}_{m,n,p}^{\mathrm{a,mp,tC}}$, $(1 \leq i \leq N)$ are such that $\|A_i\|_2 < 1 (1 \leq i \leq N)$, then by construction the 2-norm of the A matrix of the average LDS will also be less than 1. Hence, the average LDS will be asymptotically stable. Moreover, as mentioned in Sect. 8.4.3, in the case of $\mathcal{SL}_{m,n,p}^{\mathrm{a,mp,min}}$ we may employ the subbundle of balanced realizations as the standardized subbundle. It turns out that in this case preserving stability (by construction) can be easier, but the averaging algorithm gets more involved (see [2] for some more details).

Obviously, the above alignment distance based on (8.14) is only an example. In a pattern recognition application, a large class of such distances can be constructed and among them a suitable one can be chosen or they can be combined in a machine learning framework (such distances may even correspond to different standardizations).

8.5.1 Extensions

Now, we briefly point to some possible directions along which this basic idea can be extended (see also [2]). First, note that the Frobenius norm in (8.14) can be replaced by any other unitary invariant matrix norm (e.g., the nuclear norm). A less trivial extension is to get rid of $O(m)$ in (8.15) by passing to covariance matrices. For example, in the case of $\widetilde{\mathcal{OSL}}^{\text{a,mp,tC}}_{m,n,p}$ it is easy to verify that $\mathcal{SL}^{\text{a,mp,tC}}_{m,n,p} = \widetilde{\mathcal{OSL}}^{\text{a,mp,tC,cv}}_{m,n,p}/(O(n)\times I_m)$, where $\widetilde{\mathcal{OSL}}^{\text{a,mp,tC,cv}}_{m,n,p} = \{(A,Z,C,S)|(A,B,C,D)\in \widetilde{\mathcal{OSL}}^{\text{a,mp,tC}}_{m,n,p}, Z = BB^\top, S = DD^\top\}$. On this standardized subspace one only has the action of $O(n)$ which we denote as $Q\star(A,Z,C,S) = (Q^\top AQ, Q^\top ZQ, CQ, S)$. One can use the same ambient distance on this space as in (8.14) and get

$$d^2_{\Sigma_{m,n,p}}(M_1,M_2) = \min_{Q\in O(n)} \tilde{d}^2_{\widetilde{\mathcal{O}\Sigma}_{m,n,p}}\big(Q\star R_1, R_2\big), \tag{8.17}$$

for realizations $R_1, R_2 \in \widetilde{\mathcal{OSL}}^{\text{a,mp,tC,cv}}_{m,n,p}$. One could also replace the $\|\cdot\|_F$ in the terms associated with B and D in (8.14) with some known distances in the spaces of positive definite matrices or positive-semi-definite matrices of fixed rank (see e.g., [14, 63]). Another possible extension is, e.g., to consider other submanifolds of $\widetilde{\mathcal{OSL}}^{\text{a,mp,tC}}_{m,n,p}$, e.g., a submanifold where $\|C\|_F = \|B\|_F = 1$. In this case the corresponding alignment distance is essentially a scale invariant distance, i.e., two processes which are scaled version of one another will have zero distance. A more significant and subtle extension is to extend the underlying space of LDSs of fixed size and order n to that of fixed size but (minimal) order not larger than n. The details of this approach will appear a later work.

8.6 Conclusion

In this paper our focus was the geometrization of spaces of stochastic processes generated by LDSs of fixed size and order, for use in pattern recognition of high-dimensional time-series data (e.g., in the prototype Problem 1). We reviewed some of the existing approaches. We then studied the newly developed class of group action induced distances called the alignment distances. The approach is a general and flexible geometrization framework, based on the quotient structure of the space of such LDSs, which leads to a large class of extrinsic distances. The theory of alignment distances and their properties is still in early stages of development and we are hopeful to be able to tackle some interesting problems in control theory as well as pattern recognition in time-series data.

318735_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:27/3/2014 Pages: 24 Layout: T1-Standard

Acknowledgments The authors are thankful to the anonymous reviewers for their insightful comments and suggestions, which helped to improve the quality of this paper. The authors also thank the organizers of the GSI 2013 conference and the editor of this book Prof. Frank Nielsen. This work was supported by the Sloan Foundation and by grants ONR N00014-09-10084, NSF 0941362, NSF 0941463, NSF 0931805, and NSF 1335035.

AQ1

References

1. Afsari, B., Chaudhry, R., Ravichandran, A., Vidal, R.: Group action induced distances for averaging and clustering linear dynamical systems with applications to the analysis of dynamic visual scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
2. Afsari, B., Vidal, R.: The alignment distance on spaces of linear dynamical systems. In: IEEE Conference on Decision and Control (2013)
3. Afsari, B., Vidal, R.: Group action induced distances on spaces of high-dimensional linear stochastic processes. In: Geometric Science of Information, LNCS, vol. 8085, pp. 425–432 (2013)
4. Amari, S.I.: Differential geometry of a parametric family of invertible linear systems-Riemannian metric, dual affine connections, and divergence. Math. Syst. Theory **20**, 53–82 (1987)
5. Amari, S.I., Nagaoka, H.: Methods of information geometry. In: Translations of Mathematical Monographs, vol. 191. American Mathematical Society, Providence (2000)
6. Anderson, B.D., Deistler, M.: Properties of zero-free spectral matrices. IEEE Trans. Autom. Control **54**(10), 2365–5 (2009)
7. Aoki, M.: State Space Modeling of Time Series. Springer, Berlin (1987)
8. Barbaresco, F.: Information geometry of covariance matrix: Cartan-Siegel homogeneous bounded domains, Mostow/Berger fibration and Frechet median. In: Matrix Information Geometry, pp. 199–255. Springer, Berlin (2013)
9. Basseville, M.: Distance measures for signal processing and pattern recognition. Sig. Process. **18**, 349–9 (1989)
10. Basseville, M.: Divergence measures for statistical data processingan annotated bibliography. Sig. Process. **93**(4), 621–33 (2013)
11. Bauer, D., Deistler, M.: Balanced canonical forms for system identification. IEEE Trans. Autom. Control **44**(6), 1118–1131 (1999)
12. Béjar, B., Zappella, L., Vidal, R.: Surgical gesture classification from video data. In: Medical Image Computing and Computer Assisted Intervention, pp. 34–41 (2012)
13. Boets, J., Cock, K.D., Moor, B.D.: A mutual information based distance for multivariate Gaussian processes. In: Modeling, Estimation and Control, Festschrift in Honor of Giorgio Picci on the Occasion of his Sixty-Fifth Birthday, Lecture Notes in Control and Information Sciences, vol. 364, pp. 15–33. Springer, Berlin (2007)
14. Bonnabel, S., Collard, A., Sepulchre, R.: Rank-preserving geometric means of positive semi-definite matrices. Linear Algebra. Its Appl. **438**, 3202–16 (2013)
15. Byrnes, C.I., Hurt, N.: On the moduli of linear dynamical systems. In: Advances in Mathematical Studies in Analysis, vol. 4, pp. 83–122. Academic Press, New York (1979)
16. Chaudhry, R., Vidal, R.: Recognition of visual dynamical processes: Theory, kernels and experimental evaluation. Technical Report 09–01. Department of Computer Science, Johns Hopkins University (2009)
17. Chavel, I.: Riemannian Geometry: A Modern Introduction, vol. 98, 2nd edn. Cambridge University Press, Cambridge (2006)
18. Cock, K.D., Moor, B.D.: Subspace angles and distances between ARMA models. Syst. Control Lett. **46**(4), 265–70 (2002)

19. Corduas, M., Piccolo, D.: Time series clustering and classification by the autoregressive metric. Comput. Stat. Data Anal. **52**(4), 1860–72 (2008)
20. Deistler, M., Anderson, B.O., Filler, A., Zinner, C., Chen, W.: Generalized linear dynamic factor models: an approach via singular autoregressions. Eur. J. Control **3**, 211–24 (2010)
21. Devroye, L.: A probabilistic Theory of Pattern Recognition, vol. 31. Springer, Berlin (1996)
22. Doretto, G., Chiuso, A., Wu, Y., Soatto, S.: Dynamic textures. Int. J. Comput. Vision **51**(2), 91–109 (2003)
23. Ferrante, A., Pavon, M., Ramponi, F.: Hellinger versus Kullback-Leibler multivariable spectrum approximation. IEEE Trans. Autom. Control **53**(4), 954–67 (2008)
24. Forni, M., Hallin, M., Lippi, M., Reichlin, L.: The generalized dynamic-factor model: Identification and estimation. Rev. Econ. Stat. **82**(4), 540–54 (2000)
25. Georgiou, T.T., Karlsson, J., Takyar, M.S.: Metrics for power spectra: an axiomatic approach. IEEE Trans. Signal Process. **57**(3), 859–67 (2009)
26. Gray, R., Buzo, A., Gray Jr, A., Matsuyama, Y.: Distortion measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 367–76 (1980)
27. Gray, R.M.: Probability, Random Processes, and Ergodic Properties. Springer, Berlin (2009)
28. Gray, R.M., Neuhoff, D.L., Shields, P.C.: A generalization of Ornstein's $\bar{d}$ distance with applications to information theory. The Ann. Probab. **3**, 315–328 (1975)
29. Gray Jr, A., Markel, J.: Distance measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. **24**(5), 380–91 (1976)
30. Grenander, U.: Abstract Inference. Wiley, New York (1981)
31. Hannan, E.J.: Multiple Time Series, vol. 38. Wiley, New York (1970)
32. Hannan, E.J., Deistler, M.: The Statistical Theory of Linear Systems. Wiley, New York (1987)
33. Hanzon, B.: Identifiability, Recursive Identification and Spaces of Linear Dynamical Systems, vol. 63–64. Centrum voor Wiskunde en Informatica (CWI), Amsterdam (1989)
34. Hanzon, B., Marcus, S.I.: Riemannian metrics on spaces of stable linear systems, with applications to identification. In: IEEE Conference on Decision & Control, pp. 1119–1124 (1982)
35. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer, New York (2003)
36. Hazewinkel, M.: Moduli and canonical forms for linear dynamical systems II: the topological case. Math. Syst. Theory **10**, 363–85 (1977)
37. Helmke, U.: Balanced realizations for linear systems: a variational approach. SIAM J. Control Optim. **31**(1), 1–15 (1993)
38. Jiang, X., Ning, L., Georgiou, T.T.: Distances and Riemannian metrics for multivariate spectral densities. IEEE Trans. Autom. Control **57**(7), 1723–35 (2012)
39. Jimenez, N.D., Afsari, B., Vidal, R.: Fast Jacobi-type algorithm for computing distances between linear dynamical systems. In: European Control Conference (2013)
40. Kailath, T.: Linear Systems. Prentice Hall, NJ (1980)
41. Katayama, T.: Subspace Methods for System Identification. Springer, Berlin (2005)
42. Kazakos, D., Papantoni-Kazakos, P.: Spectral distance measures between Gaussian processes. IEEE Trans. Autom. Control **25**(5), 950–9 (1980)
43. Kendall, D.G., Barden, D., Carne, T.K., Le, H.: Shape and Shape Theory. Wiley Series In Probability And Statistics. Wiley, New York (1999)
44. Kobayashi, S., Nomizu, K.: Foundations of Differential Geometry Volume I. Wiley Classics Library Edition. Wiley, New York (1963)
45. Krishnaprasad, P.S.: Geometry of Minimal Systems and the Identification Problem. PhD thesis, Harvard University (1977)
46. Krishnaprasad, P.S., Martin, C.F.: On families of systems and deformations. Int. J. Control **38**(5), 1055–79 (1983)
47. Lee, J.M.: Introduction to Smooth Manifolds. Springer, Graduate Texts in Mathematics (2002)
48. Liao, T.W.: Clustering time series data—a survey. Pattern Recogn. **38**, 1857–74 (2005)
49. Makhoul, J.: Linear prediction: a tutorial review. Proc. IEEE **63**(4), 561–80 (1975)
50. Martin, A.: A metric for ARMA processes. IEEE Trans. Signal Process. **48**(4), 1164–70 (2000)

51. Moor, B.D., Overschee, P.V., Suykens, J.: Subspace algorithms for system identification and stochastic realization. Technical Report ESAT-SISTA Report 1990–28, Katholieke Universiteit Leuven (1990)
52. Moore, B.C.: Principal component analysis in linear systems: Controllability, observability, and model reduction. IEEE Trans. Autom. Control **26**, 17–32 (1981)
53. Ning, L., Georgiou, T.T., Tannenbaum, A.: Matrix-valued Monge-Kantorovich optimal mass transport. arXiv, preprint arXiv:1304.3931 (2013)
54. Nocerino, N., Soong, F.K., Rabiner, L.R., Klatt, D.H.: Comparative study of several distortion measures for speech recognition. Speech Commun. **4**(4), 317–31 (1985)
55. Ober, R.J.: Balanced realizations: canonical form, parametrization, model reduction. Int. J. Control **46**(2), 643–70 (1987)
56. Papoulis, A., Pillai, S.U.: Probability, random variables and stochastic processes with errata sheet. McGraw-Hill Education, New York (2002)
57. Piccolo, D.: A distance measure for classifying ARIMA models. J. Time Ser. Anal. **11**(2), 153–64 (1990)
58. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice-Hall International, NJ (1993)
59. Rao, M.M.: Stochastic Processes: Inference Theory, vol. 508. Springer, New York (2000)
60. Ravichandran, A., Vidal, R.: Video registration using dynamic textures. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 158–171 (2011)
61. Ravishanker, N., Melnick, E.L., Tsai, C.-L.: Differential geometry of ARMA models. J. Time Ser. Anal. **11**(3), 259–274 (1990)
62. Rozanov, Y.A.: Stationary Random Processes. Holden-Day, San Francisco (1967)
63. Vandereycken, B., Absil, P.-A., Vandewalle, S.: A Riemannian geometry with complete geodesics for the set of positive semi-definite matrices of fixed rank. Technical Report Report TW572, Katholieke Universiteit Leuven (2010)
64. Vishwanathan, S., Smola, A., Vidal, R.: Binet-Cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. Int. J. Comput. Vision **73**(1), 95–119 (2007)
65. Youla, D.: On the factorization of rational matrices. IRE Trans. Inf. Theory **7**(3), 172–189 (1961)
66. Younes, L.: Shapes and Diffeomorphisms. In: Applied Mathematical Sciences, vol. 171. Springer, New York (2010)

318735_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:**27/3/2014** Pages: **24** Layout: **T1-Standard**