A Mutual Information Based Distance for Multivariate Gaussian Processes*

Jeroen Boets, Katrien De Cock, and Bart De Moor**

K.U.Leuven, Dept. of Electrical Engineering (ESAT-SCD)
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{jeroen.boets,katrien.decock,bart.demoor}@esat.kuleuven.be

Dedicated to Giorgio Picci on the occasion of his 65th birthday.

Summary. In this paper a new distance on the set of multivariate Gaussian linear stochastic processes is proposed based on the notion of mutual information. The definition of the distance is inspired by various properties of the mutual information of past and future of a stochastic process. For two special classes of stochastic processes this mutual information distance is shown to be equal to a cepstral distance. For general multivariate processes, the behavior of the mutual information distance is similar to the behavior of an *ad hoc* defined multivariate cepstral distance.

1 Introduction

This paper is concerned with realization and identification of linear stochastic processes, topics that are central in Giorgio Picci's research interests. With his work in the last decennia he is one of the great inspirators for the development of subspace identification for stochastic processes, to which he also contributed several papers [24, 27]. Within our research group quite some work was done in subspace identification in the nineties [33, 34]. Through this way, Giorgio, we would like to thank you for the countless interesting insights you shared with us and other researchers, but especially for your great friendship. Ad multos annos!

* Research supported by Research Council KUL: GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering (OPTEC), several PhD/postdoc & fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects, G.0407.02 (support vector machines), G.0197.02 (power islands), G.0141.03 (Identification and cryptography), G.0491.03 (control for intensive care glycemia), G.0120.03 (QIT), G.0452.04 (new quantum algorithms), G.0499.04 (Statistics), G.0211.05 (Nonlinear), G.0226.06 (cooperative systems and optimization), G.0321.06 (Tensors), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, McKnow-E, Eureka-Flite2; Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011) ; EU: ERNSI.

** Jeroen Boets is a research assistant with the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) at the K.U.Leuven, Belgium. Dr. Katrien De Cock is a postdoctoral researcher at the K.U.Leuven, Belgium. Prof. Dr. Bart De Moor is a full professor at the K.U.Leuven, Belgium.

A. Chiuso et al. (Eds.): Modeling, Estimation and Control, LNCIS 364, pp. 15–33, 2007.
springerlink.com

In some of our recent work [8, 9] we have established a nice framework with interesting relations between notions from three different disciplines: system theory, information theory and signal processing. These relations are illustrated in a schematic way in Figure 1. The processes considered in the framework are scalar Gaussian linear time-invariant (LTI) stochastic processes. Centrally located in Figure 1 are the principal angles and their statistical counterparts, the canonical correlations. These notions will be explained in Section 3. Through a first link in the figure, expressions are obtained for the mutual information of past and future of a process as a function of its model parameters, by computing the canonical correlations between past and future of the process. Secondly, the notion of subspace angles between two stochastic processes allows to find new expressions for an existing cepstral distance as a function of the model description of the processes. And finally, the definition of a distance between scalar stochastic processes based on mutual information was proven to result in exactly this same cepstral distance.

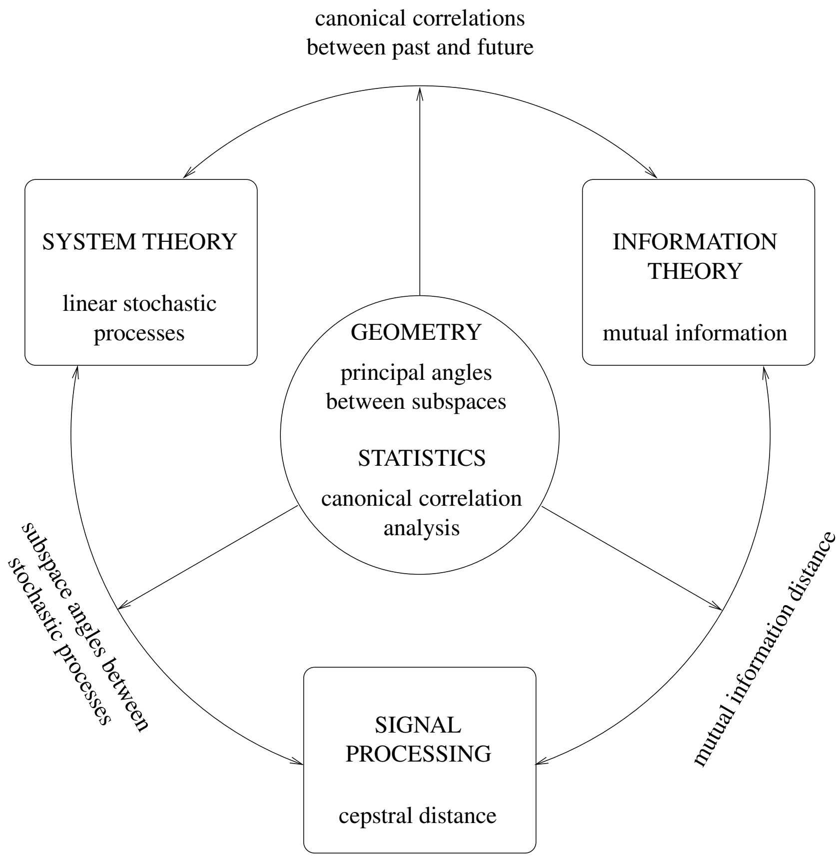

Fig. 1. A schematic representation of the relations between system theory, information theory and signal processing for scalar stochastic processes

In this paper we wish to give a start to the extension of the framework in Figure 1 to multivariate processes. We mainly focus on one aspect of the figure, namely the mutual information distance. More specifically, we define in this paper a new mutual information based distance on the set of *multivariate* Gaussian LTI stochastic processes.

The idea of defining a distance for this kind of processes is not new. Many distances have been considered in the past, both for scalar and multivariate processes. Specifically for scalar processes a lot of distances are defined directly on the basis of the power spectrum, the log-power spectrum or the power cepstrum of the processes [3,13,14,18]. A difficulty with these distances is that some of them can not be generalized in a trivial manner to multivariate processes. Cepstral distances for instance in their definition involve some definition of the logarithm of the power spectrum of the processes.

Several of the distances defined for both scalar and multivariate stochastic processes are based on information-theoretic measures. By considering a stochastic process as an infinite-dimensional random variable, one can define e.g. the (asymptotic) Kullback-Leibler (K-L) divergence, Chernoff divergence and Bhattacharyya divergence of two processes [22,25,29,30,31,32]. Often, the processes are assumed to be Gaussian, in which case computationally tractable formulas can be derived.

Mutual information is an information-theoretic measure too. However, it is not applicable in the same sense as the above measures. The difference is that the mutual information of two random variables does not measure the similarity (or dissimilarity) of their probability densities. Instead it is a measure for the *dependence* of two random variables. Since the goal in this paper is to achieve a distance on the set of stochastic processes (without assuming information on their mutual dependencies), several intermediate steps must be taken. These steps are explained in the paper and are inspired by previous work in [6,8,9] (see Figure 1).

Distances between stochastic processes or time series have been used in many different areas. Among the most common are speech recognition [3,13,14], biomedical applications [2,12,23] and video processing [4,11]. The distances are typically applied in a clustering or classification context.

The paper is organized as follows. In Section 2 we describe the model class we work with: Gaussian LTI stochastic dynamical models. Section 3 recalls the notions of principal angles between two subspaces, canonical correlations and mutual information of two random variables, and applies these notions in the context of stochastic processes. In Section 4 a new distance between multivariate Gaussian processes is proposed based on the notion of mutual information, and its properties are investigated. Section 5 shows several additional relations that hold in the case of scalar processes. In Section 6 we investigate whether the newly defined distance admits a *cepstral nature* by defining an ad hoc power cepstrum and cepstral distance for *multivariate* stochastic processes. Section 7 states the conclusions of the paper and some remaining open problems.

2 Model Class

In this paper we consider stochastic processes $y = \{y(k)\}_{k\in\mathbb{Z}}$ whose first and second order statistics can be described by the following state space equations:

$$\begin{cases} x(k+1) = Ax(k) + Bu(k) , \\ \quad y(k) = Cx(k) + Du(k) , \end{cases} \tag{1}$$

$$E\{u(k)\} = 0 , \quad E\left\{u(k)u^\top(l)\right\} = I_p\delta_{kl} . \tag{2}$$

with I_p the identity matrix of dimension p and δ_{kl} the Kronecker delta, being 1 for $k = l$ and 0 otherwise. The variable $y(k) \in \mathbb{R}^p$ is the value of the process at time k and is called the output of the model (1)-(2). The state process $\{x(k)\}_{k\in\mathbb{Z}} \in \mathbb{R}^n$ is assumed to be stationary, which implies that A is a stable matrix (all of its eigenvalues lie strictly inside the unit circle). The unobserved input process $\{u(k)\}_{k\in\mathbb{Z}} \in \mathbb{R}^p$ is a stationary and ergodic (normalized) white noise process. Both x and u are auxiliary processes used to describe the process y in this representation. The matrix $D \in \mathbb{R}^{p\times p}$ is assumed to be of full rank. We assume throughout this paper that u and consequently also y is a *Gaussian* process. This means that the process y is fully described by (1)-(2).

The infinite controllability and observability matrix of the model (1) are defined as:

$$\begin{aligned} \mathcal{C} &= \left(B \ AB \ A^2B \cdots\right) , \\ \Gamma &= \left(C^\top \ (CA)^\top \ (CA^2)^\top \cdots\right)^\top , \end{aligned}$$

respectively. The model (1) is assumed to be minimal, meaning that $\mathcal{C}$ and Γ are of full rank n. The Gramians corresponding to $\mathcal{C}$ and Γ are the unique and positive definite solution of the controllability and observability Lyapunov equation, respectively:

$$\begin{aligned} \mathcal{C}\mathcal{C}^\top = P = APA^\top + BB^\top , \\ \Gamma^\top\Gamma = Q = A^\top QA + C^\top C . \end{aligned} \tag{3}$$

The controllability Gramian P is also equal to the state covariance matrix, i.e. $P = E\left\{x(k)x^\top(k)\right\}$.

The model (1) is further assumed to be minimum-phase, meaning that its zeros (eigenvalues of $A - BD^{-1}C$) lie strictly inside the unit circle. The inverse model can then be derived from (1) by rewriting it as

$$\begin{cases} x(k+1) = (A - BD^{-1}C)x(k) + BD^{-1}y(k) , \\ \quad u(k) = \qquad -D^{-1}Cx(k) + \quad D^{-1}y(k) , \end{cases} \tag{4}$$

and is denoted with a subscript $(\cdot)_z$:

$$(A_z, B_z, C_z, D_z) = (A - BD^{-1}C, BD^{-1}, -D^{-1}C, D^{-1}).$$

Analogously, the controllability and observability matrices and Gramians of the inverse model (4) are denoted by $\mathcal{C}_z$, Γ_z, P_z and Q_z. The matrix Q_z, for instance, is the solution of

$$Q_z = (A - BD^{-1}C)^\top Q_z(A - BD^{-1}C) + C^\top D^{-\top}D^{-1}C . \tag{5}$$

Along with the descriptions (1) and (4), a transfer function can be defined from u to y and from y to u, respectively:

$$\begin{aligned} h(z) &= C(zI - A)^{-1}B + D , \\ h^{-1}(z) &= -D^{-1}C(zI - (A - BD^{-1}C))^{-1}BD^{-1} + D^{-1} . \end{aligned} \tag{6}$$

Modulo a similarity transformation of the state space model (A, B, C, D) into $(T^{-1}AT, T^{-1}B, CT, D)$ with nonsingular T, there is a one-to-one correspondence between the descriptions (1) and (6). From each of both descriptions, augmented with (2), the second order statistics of the process y can be derived, i.e. its autocovariance sequence

$$\Lambda(s) = E\left\{y(k)y^\top(k-s)\right\} = \begin{cases} CPC^\top + DD^\top & s = 0\,, \\ CA^{s-1}G & s > 0\,, \\ G^\top (A^\top)^{|s|-1} C^\top & s < 0\,, \end{cases} \tag{7}$$

with $G = E\left\{x(k+1)y^\top(k)\right\} = APC^\top + BD^\top$, or equivalently its spectral density function

$$\Phi(z) = \sum_{s=-\infty}^{+\infty} \Lambda(s) z^{-s} = h(z)h^\top(z^{-1})\,. \tag{8}$$

As stated before, Gaussian processes (which we assume) are fully described by their first and second order statistical properties. Therefore a zero-mean process $\{y(k)\}_{k\in\mathbb{Z}}$ is also fully described by (7) or (8). From equation (8) it can thus be seen that $h(z)$ is not uniquely defined for the process y since the transfer functions $h(z)$ and $h(z)V$ with V a unitary $p \times p$ matrix correspond to the same spectral density function $\Phi(z)$. This is the only non-uniqueness in $h(z)$ under the given assumptions and must be kept in mind while we denote a process in this paper by *one of its* foursomes (A, B, C, D) or *one of its* transfer functions $h(z)$.

We also define doubly infinite block Hankel matrices of data:

$$Y = \left(\begin{array}{cccc} \vdots & \vdots & \vdots & \iddots \\ y(-2) & y(-1) & y(0) & \cdots \\ y(-1) & y(0) & y(1) & \cdots \\ \hline y(0) & y(1) & y(2) & \cdots \\ y(1) & y(2) & y(3) & \cdots \\ \vdots & \vdots & \vdots & \end{array}\right) = \left(\frac{Y_p}{Y_f}\right)\,, \tag{9}$$

corresponding to the processes $y = \{y(k)\}_{k\in\mathbb{Z}}$, $y_p = \{y(-k)\}_{k\in\mathbb{N}_0}$ and $y_f = \{y(k)\}_{k\in\mathbb{N}}$, where the subscript p stands for 'past' and f for 'future'. The block Hankel matrices U, U_p and U_f are analogously defined for the processes u, u_p and u_f.

3 Principal Angles, Canonical Correlations and Mutual Information

In this section the definitions of principal angles between two subspaces, canonical correlations of two random variables and their mutual information are recalled in Sections 3.1, 3.2 and 3.3 respectively. In Section 3.4 these notions are applied in the context of the stochastic processes defined in the previous section. Attention is drawn in particular to the mutual information of past and future of the output process y.

3.1 Principal Angles and Directions

The principal angles between two subspaces [21] are a generalization of the angle between two vectors. Suppose we are given two linear subspaces S_1 and S_2 of the ambient vector space $\mathbb{R}^n$ of dimension $d_1 < n$ and $d_2 < n$, respectively. A natural extension of the one-dimensional case is to choose a unit vector u_1 from S_1 and a unit vector v_1 from S_2 such that the angle between u_1 and v_1 is minimized. The vectors u_1 and v_1 so obtained, are called the first principal directions and the angle between them is the first principal angle θ_1. Next, choose a unit vector $u_2 \in S_1$ orthogonal to u_1 and $v_2 \in S_2$ orthogonal to v_1 such that the angle θ_2 between them is minimized. This is the second principal angle and u_2 and v_2 are the corresponding principal directions. Continue in this way until $\min(d_1, d_2)$ angles and corresponding principal vectors have been found. This informal description is now formalized.

Definition 1. Principal angles and directions
The principal angles $0 \leq \theta_1 \leq \theta_2 \leq \ldots \theta_{\min(d_1,d_2)} \leq \pi/2$ *between the subspaces* S_1 *and* S_2 *of the ambient space* $\mathbb{R}^n$ *of dimension* d_1 *and* d_2*, respectively, and the corresponding principal directions* $u_i \in S_1$ *and* $v_i \in S_2$ *are defined recursively as*

$$
\begin{aligned}
\cos\theta_1 &= \max_{\substack{u\in S_1\\ v\in S_2}} u^\top v = u_1^\top v_1 \ , \\
\cos\theta_k &= \max_{\substack{u\in S_1\\ v\in S_2}} u^\top v = u_k^\top v_k \ , \ \textit{for } k = 2, \ldots, \min(d_1, d_2) \ ,
\end{aligned}
$$

subject to $\|u\| = \|v\| = 1$ *and for* $k > 1$*:* $u^\top u_i = 0$ *and* $v^\top v_i = 0$*, where* $i = 1, \ldots, k-1$.

Let $A \in \mathbb{R}^{p\times n}$ be of rank d_1 and $B \in \mathbb{R}^{q\times n}$ of rank d_2. Then, the ordered set of $\min(d_1, d_2)$ principal angles between the row spaces of A and B is denoted by

$$
\left(\theta_1, \theta_2, \ldots, \theta_{\min(d_1,d_2)}\right) = [A \sphericalangle B] \ .
$$

In case A and B are of full row rank with $p \leq q$, the squared cosines of the principal angles between $\mathrm{row}(A)$ and $\mathrm{row}(B)$ are equal to the eigenvalues of $(AA^\top)^{-1}AB^\top (BB^\top)^{-1}BA^\top$:

$$
\cos^2 [A \sphericalangle B] = \lambda \left((AA^\top)^{-1}AB^\top(BB^\top)^{-1}BA^\top\right) \ . \tag{10}
$$

3.2 Canonical Correlations

In canonical correlation analysis [16] the interrelation of two sets of random variables is studied. It is the statistical interpretation of the geometric tool of principal angles between and principal directions in linear subspaces. The aim is to find two bases of random variables, one in each set, that are internally uncorrelated but that have maximal correlations between the two sets. The resulting basis variables are called the canonical variates and the correlation coefficients between the canonical variates are the canonical correlations.

Let V be a zero-mean p-component and W a zero-mean q-component real random variable with joint covariance matrix $Q = E\left\{\begin{pmatrix} V \\ W \end{pmatrix}\begin{pmatrix} V^\top & W^\top \end{pmatrix}\right\} = \begin{pmatrix} Q_v & Q_{vw} \\ Q_{wv} & Q_w \end{pmatrix}$. In case Q_v and Q_w are full rank matrices, and $p \leq q$, the p squared canonical correlations of V and W, which we denote by $\mathrm{cc}^2(V, W)$, can be obtained as the eigenvalues of $Q_v^{-1} Q_{vw} Q_w^{-1} Q_{wv}$:

$$\mathrm{cc}^2(V, W) = \lambda(Q_v^{-1} Q_{vw} Q_w^{-1} Q_{wv}) \,. \tag{11}$$

3.3 Mutual Information

Let V be a zero-mean p-component and W a zero-mean q-component random variable. If V and W are mutually dependent, then observing W reduces the uncertainty (or entropy) in V. Otherwise formulated, we gain information about V by observing W. Thus, the variable W must contain information about V. For the same reason V must also contain information about W. Both amounts of information are equal and are quantified as the *mutual information of* V *and* W, denoted by $I(V; W)$.

Definition 2. The mutual information of two continuous random variables [7]
Let V and W be random variables with joint probability density function $f(v, w)$ and marginal densities $f_V(v)$ and $f_W(w)$, respectively. Then, the mutual information of V and W is defined as

$$I(V; W) = \iint f(v, w) \log \frac{f(v, w)}{f_V(v) f_W(w)} \, dv \, dw \,,$$

if the integral exists.

In case of two zero-mean jointly Gaussian random variables and denoting the covariance matrix of $\begin{pmatrix} V \\ W \end{pmatrix}$ by $Q = \begin{pmatrix} Q_v & Q_{vw} \\ Q_{wv} & Q_w \end{pmatrix}$, this expression can be rewritten as

$$I(V; W) = -\frac{1}{2} \log \frac{\det Q}{\det Q_v \det Q_w} \,,$$

under the assumption that Q_v and Q_w are of full rank. In this case $I(V; W)$ is also related to the canonical correlations of V and W, here denoted by σ_k ($k = 1, \ldots, \min(p, q)$), as can be derived using equation (11):

$$I(V; W) = -\frac{1}{2} \log \prod_{k=1}^{\min(p,q)} (1 - \sigma_k^2) \,. \tag{12}$$

3.4 Application to Stochastic Processes

In this section we apply the notions defined in the previous sections to the stochastic processes y_p, y_f, u_p and u_f. A stochastic process, e.g. $\{y(k)\}_{k \in \mathbb{Z}}$, can be seen as an infinite-dimensional random variable consisting of the (ordered) concatenation of the random variables $\ldots,\ y(-2),\ y(-1),\ y(0),\ y(1),\ \ldots$ We can thus associate with the process y the random variable

$$\mathcal{Y} = \begin{pmatrix} \vdots \\ y(-2) \\ y(-1) \\ \hline y(0) \\ y(1) \\ \vdots \end{pmatrix} = \begin{pmatrix} \mathcal{Y}_p \\ \hline \mathcal{Y}_f \end{pmatrix} ,$$

$\mathcal{Y}_p$ and $\mathcal{Y}_f$ being associated with the processes y_p and y_f, and analogously $\mathcal{U}$, $\mathcal{U}_p$ and $\mathcal{U}_f$ for the processes u, u_p and u_f. This way we can compute the canonical correlations and the mutual information for any pair of these processes.

Canonical Correlations

Since we are dealing with stationary and ergodic zero-mean processes, it is readily seen from equations (10) and (11) that the canonical correlations between any two of the processes u, u_p, u_f, y, y_p and y_f are equal to the cosines of the principal angles between the row spaces of the corresponding block Hankel matrices defined in (9), e.g.:

$$\mathrm{cc}(\mathcal{U}_f, \mathcal{Y}_f) = \cos\left([U_f \sphericalangle Y_f]\right) \ . \tag{13}$$

In [8, Chap. 3] the canonical correlations of each pair of these processes were computed. Formulas were derived for the canonical correlations between the past and future output process:

$$\mathrm{cc}^2(\mathcal{Y}_p, \mathcal{Y}_f) = \lambda\left(P(Q_z^{-1} + P)^{-1}\right), 0, 0, \ldots\ ,$$

as well as for the canonical correlations between u_f and y_f:

$$\mathrm{cc}^2(\mathcal{U}_f, \mathcal{Y}_f) = \lambda\left((I_n + Q_z P)^{-1}\right), 1, 1, \ldots\ ,$$

where P and Q_z each follow from a Lyapunov equation (see (3)-(5)). We denote the non-trivial correlations of y_p and y_f by ρ_k, and those of u_f and y_f by τ_k, as follows:

$$\begin{aligned} \rho_k^2 &= \lambda\left(P(Q_z^{-1} + P)^{-1}\right) \quad (k = 1, \ldots, n)\ , \\ \tau_k^2 &= \lambda\left((I_n + Q_z P)^{-1}\right) \quad (k = 1, \ldots, n)\ . \end{aligned} \tag{14}$$

It can be shown that $\rho_k^2 + \tau_k^2 = 1$, for $k = 1, \ldots, n$. These results together with the canonical correlations of the other pairs of processes are summarized in Table 1.

Mutual Information of Past and Future of a Process

Using the relation (12) for Gaussian processes, we can compute from Table 1 the mutual information of each pair of processes. A pair of processes that has at least one canonical correlation equal to 1 does not have a finite amount of mutual information.

Table 1. Overview of the canonical correlations of each pair of processes, where k goes from 1 to n

	$\mathcal{U}_p$	$\mathcal{Y}_p$	$\mathcal{U}_f$	$\mathcal{Y}_f$
$\mathcal{U}_p$	$1, 1, \ldots$	$1, 1, \ldots$	$0, 0, \ldots$	$\rho_k, 0, 0, \ldots$
$\mathcal{Y}_p$	$1, 1, \ldots$	$1, 1, \ldots$	$0, 0, \ldots$	$\rho_k, 0, 0, \ldots$
$\mathcal{U}_f$	$0, 0, \ldots$	$0, 0, \ldots$	$1, 1, \ldots$	$\sqrt{1-\rho_k^2}, 1, 1, \ldots$
$\mathcal{Y}_f$	$\rho_k, 0, 0, \ldots$	$\rho_k, 0, 0, \ldots$	$\sqrt{1-\rho_k^2}, 1, 1, \ldots$	$1, 1, \ldots$

Looking at relation (13) between canonical correlations and principal angles we can say that these processes intersect, since they have a principal angle equal to zero. Conversely, processes that are orthogonal to each other (all canonical correlations equal to 0 or all principal angles equal to $\pi/2$) have mutual information equal to zero. This is for instance the case for u_p and u_f, past and future of the white noise process u. However, processing this white noise u through the filter $h(z)$ (in general) introduces a time correlation in the resulting process y, which appears as a certain amount of mutual information between its past y_p and future y_f, denoted interchangeably by I_{pf}, $I_{\mathrm{pf}}\{y\}$ or $I_{\mathrm{pf}}\{h(z)\}$:

$$I_{\mathrm{pf}} = I(y_p; y_f) = -\frac{1}{2}\log\prod_{k=1}^{n}(1-\rho_k^2) = -\frac{1}{2}\log\prod_{k=1}^{n}\tau_k^2 = \frac{1}{2}\log\det\left(I_n + Q_z P\right) \ . \tag{15}$$

Note that ρ_k, τ_k ($k = 1, \ldots, n$) and consequently also I_{pf} are unique for a given stochastic process, since P and Q_z do not change when $h(z)$ is right-multiplied by a unitary matrix, and a similarity transformation of the state space model does not alter the eigenvalues of the product $Q_z P$. So if we write $I_{\mathrm{pf}}\{h(z)\}$ or $\rho_k\{h(z)\}$, this must not be understood as a characteristic of the transfer function $h(z)$ but rather as a characteristic of the process y with spectral density $\Phi(z) = h(z)h^\top(z^{-1})$.

Properties of I_{pf}

The mutual information I_{pf} of past and future of a stochastic process y is the amount of information that the past provides about the future and vice versa. Through (15) it is closely connected to the canonical correlations of y_p and y_f. The problem of characterizing this dependence of past and future of a stationary process has received a great deal of attention because of its implications for the prediction theory of Gaussian processes (see [17, 19, 20]). Inspired by the use of canonical correlation analysis in stochastic realization theory [1], a stochastic model reduction technique based on the mutual information of the past and the future has been proposed by Desai and Pal [10], which is also used in stochastic subspace identification [27, 34]. Li and Xie used the past-future mutual information for model selection and order determination problems in [26]. We now state some of the properties of I_{pf}.

(a) $I_{\mathrm{pf}} = 0 \Leftrightarrow h(z) = D$ (see (1))

Since y is Gaussian, $I_{\mathrm{pf}} = 0$ is equivalent with y_p and y_f being uncorrelated, thus $\Lambda(s) = 0_p$ for $s \neq 0$. From stochastic realization theory then follows that $h(z)$ has order zero.

(b) $I_{\mathrm{pf}} \in [0, +\infty)$

This follows from relation (15) and the fact that $\rho_k \in [0, 1)$. Indeed, in [15] it is shown that the number of unit canonical correlations of y_p and y_f is equal to the number of zeros of $h(z)$ on the unit circle. Since $h(z)$ is assumed to be minimum-phase (see Section 2), this number is zero.

(c) I_{pf} (strictly) increases with each increase of a canonical correlation ρ_k ($k = 1, \ldots, n$).

This follows immediately from relation (15) and property (b).

(d) $I_{\mathrm{pf}}\{h(z)\} = I_{\mathrm{pf}}\{Th(z)\}$ for a nonsingular constant matrix $T \in \mathbb{R}^{p \times p}$.

This follows from the definition of canonical correlations or principal angles, since left-multiplying the output variables $y(k)$ ($k \in \mathbb{Z}$) with T does not change the row spaces of Y_p and Y_f. Consequently, the canonical correlations ρ_k and the mutual information I_{pf} do not change.

(e) $I_{\mathrm{pf}}\{h(z)\} = I_{\mathrm{pf}}\{h^{-\top}(z)\}$

Equation (14) shows that the past-future canonical correlations ρ_k ($k = 1, \ldots, n$) only depend on the eigenvalues of the product matrix $Q_z P$. Noting that the state space description of the transpose of the inverse model is given by $h^{-\top}(z) = (A_z^\top, C_z^\top, B_z^\top, D_z^\top)$, it can be seen from (3) that the controllability Gramian of $h^{-\top}(z)$ is given by Q_z, while the observability Gramian of its inverse model $h^\top(z) = (A^\top, C^\top, B^\top, D^\top)$ is equal to P. Consequently, the canonical correlations ρ_k and the mutual information I_{pf} are equal for the transfer functions $h(z)$ and $h^{-\top}(z)$. This invariance property does not, in general, hold for $h(z)$ and $h^{-1}(z)$ since the eigenvalues of $Q_z P$ are usually not equal to those of QP_z.

(f) For $\Phi(z) = \begin{pmatrix} \Phi_1(z) & 0_{p_1 \times p_2} \\ 0_{p_2 \times p_1} & \Phi_2(z) \end{pmatrix}$, it holds that $I_{\mathrm{pf}}\{y\} = I_{\mathrm{pf}}\{y_1\} + I_{\mathrm{pf}}\{y_2\}$.

In this case the p_1-variate process y_1 and the p_2-variate process y_2, constituting the process y, are completely uncorrelated. Therefore, the canonical correlations of y_p and y_f are on the one hand the canonical correlations between y_{1_p} and y_{1_f}, and on the other hand the canonical correlations between y_{2_p} and y_{2_f}: $\rho_k\{y\}$ ($k = 1, \ldots, n_1 + n_2$) is the union of $\rho_k\{y_1\}$ ($k = 1, \ldots, n_1$) and $\rho_k\{y_2\}$ ($k = 1, \ldots, n_2$), with n_1 and n_2 the orders of the processes y_1 and y_2. The result then follows from relation (15).

Properties (a)-(c) indicate that I_{pf} measures the amount of correlation that exists between y_p and y_f, being zero for a white noise process and increasing with each increase of a correlation ρ_k between y_p and y_f. This suggests that I_{pf} can be used as a measure for the amount of *dynamics* in the process y where dynamics are defined in terms of the correlation or the dependence that exists between all future values and all past values of the process at any time instant.

4 A Distance Between Multivariate Gaussian Processes

In this section we define a new distance between multivariate Gaussian processes based on the notion of mutual information. In Section 4.1 the distance is defined and its metric properties are investigated, while in Section 4.2 we show a way to compute the distance.

4.1 Definition and Metric Properties

We propose as a new distance on the set of multivariate Gaussian processes: the *mutual information distance*, denoted by $\mathrm{d}_{\mathrm{mi}}(y_1, y_2)$.

Definition 3. The mutual information distance between two Gaussian processes *The mutual information distance between two Gaussian linear stochastic processes y_1 and y_2 with transfer function descriptions $h_1(z)$ and $h_2(z)$ is denoted by $\mathrm{d}_{\mathrm{mi}}(y_1, y_2)$ and is defined as*

$$\mathrm{d}^2_{\mathrm{mi}}(y_1, y_2) = I_{\mathrm{pf}}\left\{h_{12}(z)\right\}, \quad \text{with } h_{12}(z) = \begin{pmatrix} h_1^{-1}(z)h_2(z) & 0_p \\ 0_p & h_2^{-1}(z)h_1(z) \end{pmatrix}.$$

The first thing to note is that the mutual information distance $\mathrm{d}_{\mathrm{mi}}(y_1, y_2)$ is a property of the processes y_1 and y_2, and not of the particular transfer functions $h_1(z)$ and $h_2(z)$. Indeed, substituting $\{h_1(z), h_2(z)\}$ by the equivalent $\{h_1(z)V_1, h_2(z)V_2\}$ with V_1, V_2 constant unitary matrices (see (8)), corresponds to left- and right-multiplying $h_{12}(z)$ by a constant unitary matrix. This has no influence on $I_{\mathrm{pf}}\left\{h_{12}\right\}$ (see property (d) in Section 3.4).

Following the discussion at the end of Section 3.4, $\mathrm{d}_{\mathrm{mi}}(y_1, y_2)$ can be interpreted as a measure for the amount of dynamics in the process y_{12} associated with the transfer function $h_{12}(z)$. It is clear that $\mathrm{d}_{\mathrm{mi}}\{y_1, y_1\} = 0$ since $h_{12}(z)$ is in that case a constant matrix and y_{12} is consequently white noise. This also clarifies why the 'ratio' of $h_1(z)$ and $h_2(z)$ is found in $h_{12}(z)$, instead of for instance the difference. From Definition 3 it is also immediately seen that $\mathrm{d}_{\mathrm{mi}}(y_1, y_2) = \mathrm{d}_{\mathrm{mi}}(g(z)y_1, g(z)y_2)$ for arbitrary transfer functions $g(z)$ satisfying the conditions stated in Section 2 (e.g. being square, stable and minimum-phase). Filtering the processes y_1 and y_2 by a common filter $g(z)$ does not change their mutual information distance.

The following properties hold for the mutual information distance:

1. $\mathrm{d}_{\mathrm{mi}}(y_1, y_2) \geq 0$
2. $\mathrm{d}_{\mathrm{mi}}(y_1, y_2) = 0 \Leftrightarrow h_2(z) = h_1(z)T$ with T a constant square nonsingular matrix. This follows from property (a) in Section 3.4.
3. $\mathrm{d}_{\mathrm{mi}}(y_1, y_2) = \mathrm{d}_{\mathrm{mi}}(y_2, y_1)$ is symmetric. This follows immediately from Definition 3.

Examples have shown that $\mathrm{d}_{\mathrm{mi}}(y_1, y_2)$ does not in general satisfy the triangle inequality[1]. The distance thus satisfies only two of the four properties of a true metric (non-negativity and symmetry). However, if we define a set of equivalence classes of stochastic processes, where two processes with transfer functions $h_1(z)$ and $h_2(z)$ are equivalent if and only if there exists a constant square nonsingular matrix T such that $h_2(z) = h_1(z)T$, then the mutual information distance $\mathrm{d}_{\mathrm{mi}}(y_1, y_2)$ defined on this set of equivalence classes, satisfies all metric properties but the triangle inequality. It is then called a *semimetric*.

4.2 Computation

From property (f) in Section 3.4 it follows that

$$\mathrm{d}^2_{\mathrm{mi}}(y_1, y_2) = I_{\mathrm{pf}}\left\{h_1^{-1}(z)h_2(z)\right\} + I_{\mathrm{pf}}\left\{h_2^{-1}(z)h_1(z)\right\}. \tag{16}$$

Using this property we now show a way to compute $\mathrm{d}_{\mathrm{mi}}(y_1, y_2)$ making use of the state space descriptions of $h_1(z)$ and $h_2(z)$ of orders n_1 and n_2 respectively. Equations (15) and (16) show that we need to compute the controllability and observability Gramians of both $h_1^{-1}(z)h_2(z)$ and $h_2^{-1}(z)h_1(z)$. This can be easily done by solving the Lyapunov equations (3) from the state space descriptions of both transfer functions. As an example we give a possible state space description of $h_1^{-1}(z)h_2(z)$ denoted by $(A_{12}, B_{12}, C_{12}, D_{12})$:

$$A_{12} = \begin{pmatrix} A_2 & 0_{n_2 \times n_1} \\ B_{z_1}C_2 & A_{z_1} \end{pmatrix}, \; B_{12} = \begin{pmatrix} B_2 \\ B_{z_1}D_2 \end{pmatrix}, \; C_{12} = \begin{pmatrix} D_{z_1}C_2 & C_{z_1} \end{pmatrix}, \; D_{12} = D_{z_1}D_2 \,,$$

with $(A_{z_1}, B_{z_1}, C_{z_1}, D_{z_1}) = (A_1 - B_1D_1^{-1}C_1, B_1D_1^{-1}, -D_1^{-1}C_1, D_1^{-1})$. The procedure concerning $h_2^{-1}(z)h_1(z)$ is analogous. Afterwards it remains to compute (16) using (15) and (3).

5 Special Case of Scalar Processes

The only relation in Figure 1 that holds for both scalar and multivariate Gaussian processes is the one between the mutual information distance and the past-future canonical correlations, which can be seen in (15). In the case of scalar processes y_1 and y_2 it follows from property (e) in Section 3.4 that (16) can be rewritten as $\mathrm{d}^2_{\mathrm{mi}}(y_1, y_2) = 2I_{\mathrm{pf}}\left\{\frac{h_1(z)}{h_2(z)}\right\} = 2I_{\mathrm{pf}}\left\{\frac{h_2(z)}{h_1(z)}\right\}$. In this case the mutual information distance is also related to so-called *subspace angles between stochastic processes* and to a cepstral distance, as was mentioned in the introduction (see Figure 1). We will shortly recall these two results in Sections 5.1 and 5.2. Based on these relations, several additional expressions for $\mathrm{d}_{\mathrm{mi}}(y_1, y_2)$ can be derived for the scalar case. For more details on this we refer to [8, Chap. 6].

[1] In the case of scalar processes or processes with diagonal spectral density function $\Phi(z)$, however, it can be shown that the triangle inequality *is* satisfied (see Sections 5.2 and 6.1 respectively).

5.1 Relation with Subspace Angles Between Scalar Stochastic Processes

Consider the situation in Figure 2 where the single-input single-output models $h_1(z)$ of order n_1 and $h_2(z)$ of order n_2 are driven by a common white noise source $\{u(k)\}_{k\in\mathbb{Z}} \in \mathbb{R}$. It can be shown that in this case only $n_1 + n_2$ canonical correlations between the future y_{1_f} and y_{2_f} of the processes y_1 and y_2 can be different from 1. If we denote these correlations by ν_k ($k = 1, \ldots, n_1 + n_2$), then the following relation was proven in [8]:

$$\mathrm{d}_{\mathrm{mi}}^2(y_1, y_2) = -\log \prod_{k=1}^{n_1+n_2} \nu_k^2 = -\log \prod_{k=1}^{n_1+n_2} \cos^2 \psi_k \ , \tag{17}$$

where the angles ψ_k ($k = 1, \ldots, n_1 + n_2$) are the $n_1 + n_2$ largest principal angles between the row spaces of the block Hankel matrices Y_{1_f} and Y_{2_f}. They are called the *subspace angles between $h_1(z)$ and $h_2(z)$*, denoted by $[h_1(z) \sphericalangle h_2(z)]$. They can be expressed as the principal angles between subspaces immediately derived from the models:

$$[h_1(z) \sphericalangle h_2(z)] = \left[\begin{pmatrix} \mathcal{C}^{(1)} \\ \mathcal{O}_z^{(2)^\top} \end{pmatrix} \sphericalangle \begin{pmatrix} \mathcal{O}_z^{(1)^\top} \\ \mathcal{C}^{(2)} \end{pmatrix} \right] . \tag{18}$$

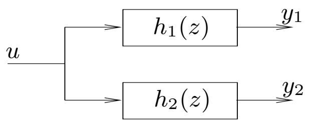

Fig. 2. Setup for the definition of subspace angles between two scalar processes

5.2 Relation with a Cepstral Distance

The power cepstrum of a scalar process y is defined as the inverse Fourier transform of the logarithm of the power spectrum of y:

$$\log \Phi(e^{j\theta}) = \sum_{k=-\infty}^{+\infty} c(k) e^{-jk\theta} \ , \tag{19}$$

where $c(k)$ is the kth cepstral coefficient of y. The sequence $\{c(k)\}_{k\in\mathbb{Z}}$ contains the same information as $\Phi(z)$ and thus also fully characterizes the zero-mean Gaussian process y. The sequence is real and even, i.e. $c(k) = c(-k)$, and can be expressed in terms of the model parameters:

$$c(k) = \begin{cases} \log D^2 & k = 0 \ , \\ \displaystyle\sum_{i=1}^{n} \frac{\alpha_i^{|k|}}{|k|} - \sum_{i=1}^{n} \frac{\beta_i^{|k|}}{|k|} & k \neq 0 \ , \end{cases} \tag{20}$$

where the poles of $h(z)$ are denoted by $\alpha_1, \ldots, \alpha_n$ and the zeros by $\beta_1, \ldots, \beta_n$. Based on the cepstral coefficients, a *weighted cepstral distance* was defined in [28]:

$$\mathrm{d}_{\mathrm{cep}}^2(y_1, y_2) = \sum_{k=0}^{+\infty} k(c_1(k) - c_2(k))^2 , \tag{21}$$

with c_1 and c_2 the cepstra of the processes y_1 and y_2 and 'cep' referring to 'cepstral'. Based on (18), this distance $\mathrm{d}_{\mathrm{cep}}$ was proven in [8, Chap. 6] (and differently also in [20]) to be equal to the mutual information distance d_{mi}, i.e.:

$$\mathrm{d}_{\mathrm{mi}}(y_1, y_2) = \mathrm{d}_{\mathrm{cep}}(y_1, y_2) . \tag{22}$$

This obviously proves that d_{mi} for scalar processes satisfies the triangle inequality. Referring to the discussion in Section 4.1 we can thus say that d_{mi} is a true metric on the set of equivalence classes of scalar stochastic processes, where two processes y_1 and y_2 are equivalent if and only if $h_2(z) = a h_1(z)$ for a non-zero real number a.

6 The Cepstral Nature of the Mutual Information Distance

The equality (22) of d_{mi} and $\mathrm{d}_{\mathrm{cep}}$ was formulated for *scalar* stochastic processes. In the case of *multivariate* processes, one would first need a definition of the power cepstrum of a multivariate process. No such definition is known to the authors of this paper. Therefore, we introduce in Section 6.1 a multivariate power cepstrum and a corresponding weighted cepstral distance, denoted by $\mathrm{d}_{\mathrm{cep}}$.

Even with this new definition, the relation (22) does not hold for general multivariate processes. However, it turnes out experimentally that d_{mi} has a *cepstral character*. This is explained in Section 6.2.

6.1 Multivariate Power Cepstrum and Cepstral Distance

No definition of the power cepstrum of a *multivariate* process y is known to the authors of this paper. Therefore, in analogy with (19), we propose to define the power cepstrum of a multivariate process y as the inverse Fourier transform of the *matrix* logarithm of the power spectrum of y:

$$\log \Phi(e^{j\theta}) = \sum_{k=-\infty}^{+\infty} c(k) e^{-jk\theta} , \tag{23}$$

where $c(k) \in \mathbb{R}^{p \times p}$ is the kth cepstral coefficient *matrix* of y. The sequence $\{c(k)\}_{k \in \mathbb{Z}}$ is real and even, and again contains the same information as $\Phi(z)$ and thus also fully characterizes the zero-mean Gaussian process y. However, no analytical expressions as in (20) are known to us for these multivariate cepstral coefficients, although in principle they could be calculated from the state space description (8) of $\Phi(z)$ by expanding the Laurent series of $\log \Phi(z)$ around the origin.

We now define in analogy with (21) a multivariate weighted cepstral distance as

$$\mathrm{d}^2_{\mathrm{cep}}(y_1, y_2) = \sum_{k=0}^{+\infty} k \|c_1(k) - c_2(k)\|^2_{\mathrm{F}} \,, \tag{24}$$

with c_1 and c_2 the cepstra of the multivariate processes y_1 and y_2, and $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm of a matrix. For scalar processes this distance coincides with the previously defined distance (21). No relation with the mutual information distance as in (22) for scalar processes holds for multivariate processes, except for diagonal $\Phi_1(z), \Phi_2(z)$ where it is easily shown that

$$\mathrm{d}^2_{\mathrm{mi}}(y_1, y_2) = \sum_{i=1}^{p} \mathrm{d}^2_{\mathrm{mi}}(y_{1,i}, y_{2,i}) = \sum_{i=1}^{p} \mathrm{d}^2_{\mathrm{cep}}(y_{1,i}, y_{2,i}) = \mathrm{d}^2_{\mathrm{cep}}(y_1, y_2) \,,$$

with $y_{1,i}$ ($i = 1, \ldots, p$) the uncorrelated scalar processes constituting y_1, and analogously for $y_{2,i}$ ($i = 1, \ldots, p$). The first equality follows from Definition 3 and property (f) in Section 3.4. The second equality follows from relation (22) for scalar processes.

The distance (24) can be computed based on the model descriptions of the processes y_1 and y_2. These allow to compute exact values of $\log \Phi(e^{j\theta})$ where θ varies over a discretization of the interval $[0, 2\pi]$. After applying the inverse fast Fourier transform (IFFT) to obtain estimates of the cepstral coefficients, one can further approximate (24) by replacing $+\infty$ in the formula by a finite L.

6.2 The Cepstral Nature of the Mutual Information Distance

For scalar processes, several simulation experiments were performed in [5] in order to compare the behavior of the cepstral distance $\mathrm{d}_{\mathrm{cep}}$, which is equal to d_{mi} because of (22), with the behavior of the $\mathbf{H}_2$ distance, denoted by $\mathrm{d}_{\mathrm{h}_2}$:

$$\mathrm{d}^2_{\mathrm{h}_2}(h_1(z), h_2(z)) = \|h_1(z) - h_2(z)\|^2_{\mathrm{h}_2} = \frac{1}{2\pi} \int_0^{2\pi} \|h_1(e^{j\theta}) - h_2(e^{j\theta})\|^2_{\mathrm{F}} d\theta \,. \tag{25}$$

In order to make $\mathrm{d}_{\mathrm{h}_2}$ a distance between processes instead of between transfer functions, we agree to fix the transfer function description of a stochastic process. We always choose the D-matrix of a model (1) or (6) to be D_{chol}, the unique Cholesky factor of $DD^\top$, which is invariant for a given stochastic process.

In this section we focus on two aspects that showed in the scalar case a difference in behavior between the cepstral distance and the $\mathbf{H}_2$ distance:

1. The influence of poles of $h_1(z)$ and $h_2(z)$ approaching the unit circle.
2. The influence of poles of $h_2(z)$ approaching the unit circle (with fixed zeros), compared to the influence of zeros of $h_2(z)$ approaching the unit circle (with fixed poles). Poles and zeros of $h_1(z)$ are kept fixed.

In order to understand why we choose these two experimental settings, one should notice an important difference between $\mathrm{d}_{\mathrm{h}_2}$ in (25) and $\mathrm{d}_{\mathrm{cep}}$ in (21) and (24), namely the presence of the *logarithm* of the power spectrum in the definition of the cepstrum (19) and (23). For the scalar case this has the following consequences:

1. High peaks in the spectrum of $h_i(z)$ (corresponding to poles close to the unit circle) have a greater influence on $\mathrm{d}_{\mathrm{h}_2}(h_1, h_2)$ than on $\mathrm{d}_{\mathrm{cep}}(h_1, h_2)$.
2. Deep valleys in the spectrum of $h_i(z)$ (corresponding to zeros close to the unit circle) have a greater influence on $\mathrm{d}_{\mathrm{cep}}(h_1, h_2)$ than on $\mathrm{d}_{\mathrm{h}_2}(h_1, h_2)$.

It can be shown that cepstral distances in the scalar case are equally dependent on the poles and zeros of $h_i(z)$: the distance between two models is equal to the distance between the inverses of the two models. The distance $\mathrm{d}_{\mathrm{h}_2}$, on the other hand, is much less sensitive to the depth of a valley than to the height of a peak in the spectrum of $h_i(z)$.

It turns out that, in the *multivariate* case, the mutual information distance d_{mi} and the cepstral distance $\mathrm{d}_{\mathrm{cep}}$ have several characteristics in common, whereas the $\mathbf{H}_2$ distance $\mathrm{d}_{\mathrm{h}_2}$ behaves very differently:

1. The distance $\mathrm{d}_{\mathrm{h}_2}(h_1, h_2)$ grows much faster than $\mathrm{d}_{\mathrm{cep}}(h_1, h_2)$ and $\mathrm{d}_{\mathrm{mi}}(h_1, h_2)$ as the poles of $h_1(z)$ and $h_2(z)$ approach the unit circle. This means that $\mathrm{d}_{\mathrm{h}_2}$ is more sensitive to high peaks in the spectrum of $h_i(z)$ than $\mathrm{d}_{\mathrm{cep}}$ and d_{mi}. The distances $\mathrm{d}_{\mathrm{cep}}$ and d_{mi} evolve quite similarly to each other.
2. The distance $\mathrm{d}_{\mathrm{h}_2}(h_1, h_2)$ grows much faster in case $h_2(z)$ has fixed zeros but poles approaching the unit circle, than in case $h_2(z)$ has fixed poles but zeros approaching the unit circle. For both the distances $\mathrm{d}_{\mathrm{cep}}(h_1, h_2)$ and $\mathrm{d}_{\mathrm{mi}}(h_1, h_2)$, on the other hand, the evolution of the distance in case of poles approaching the unit circle is very similar to the evolution in case of zeros approaching the unit circle. This means that $\mathrm{d}_{\mathrm{h}_2}$ is much more sensitive to high peaks than to deep valleys in the spectrum of $h_i(z)$, whereas $\mathrm{d}_{\mathrm{cep}}$ and d_{mi} are more or less equally sensitive. The distances $\mathrm{d}_{\mathrm{cep}}$ and d_{mi} also evolved quite similarly to each other.

With these conclusions we do not claim that one of the distances is *better* than the others. We only wish to point out some differences between them. On the basis of these differences one can choose which distance to use in a specific application.

7 Conclusions and Open Problems

7.1 Conclusions

In this paper we defined the mutual information distance on the set of multivariate Gaussian linear stochastic processes, based on the notion of mutual information of past and future of a stochastic process and inspired by the various properties of this notion. We demonstrated how it can be computed from the state space description of the processes and showed that it is a semimetric on a set of equivalence classes of stochastic processes. For two special classes of stochastic processes, namely scalar processes and processes with diagonal spectral density function, a link exists between the mutual information distance and a previously defined scalar cepstral distance.

The mutual information distance shows a behavior similar to an *ad hoc* defined multivariate cepstral distance and dissimilar from the $\mathbf{H}_2$ distance: it does not inflate when poles of the models are approaching the unit circle and it is more sensitive to differences in zeros than the $\mathbf{H}_2$ distance.

7.2 Open Problems

In this paper a possible extension for multivariate processes was considered of the theory for scalar processes described in Section 5 and Figure 1. The proposed Definition 3 of a multivariate distance however only involves the notion of mutual information and not the notions of subspace angles or cepstral distances between stochastic processes. Thus there remain quite some challenges and issues to be investigated concerning a comparable theory for multivariate stochastic processes.

Furthermore, it would be nice to have more rigorous evidence for the conclusions drawn in Section 6.2.

Multivariate Power Cepstrum and Cepstral Distance

No definition of the power cepstrum of a multivariate process is known to the authors of this paper. Therefore, we introduced an ad hoc definition (23) in Section 6.1. For these cepstral coefficients, however, no analytical expressions are known comparable to e.g. (20) for the scalar coefficients. This topic needs further investigation.

Based on the definition of a multivariate power cepstrum one can define distances in the cepstral domain. In this paper one possible approach was considered in (24) in analogy with (21). But this is clearly not the only possibility.

Subspace Angles Between Multivariate Stochastic Processes

The definition of subspace angles between scalar stochastic processes based on Figure 2 is not readily extendable to multivariate processes. The non-uniqueness of the transfer function description of a multivariate process (see the discussion below (8)) also causes non-uniqueness in the definition of the subspace angles between two multivariate processes. Further investigation is necessary to find a good way to circumvent this problem.

Relations Between System Theory, Information Theory and Signal Processing

Looking at Figure 1 for scalar processes, it is very tempting to look for similar relations in the case of multivariate processes. The two previous topics described the lack of a definition of subspace angles and cepstral distances between multivariate processes. A possible guideline in the search for these definitions could be the attempt to establish a relation with the distance d_{mi} similar to (17) and (22) for scalar stochastic processes. Alternatively, the search for definitions of subspace angles and cepstral distances between multivariate processes could also be guided by the search for a direct link between both, not necessarily through d_{mi}.

References

1. H. Akaike. Markovian representation of stochastic processes by canonical variables. *SIAM Journal on Control*, 13(1):162–173, 1975.
2. C. W. Anderson, E. A. Stolz, and S. Shamsunder. Multivariate autoregressive models for classification of spontaneous electroencephalografic signals during mental tasks. *IEEE Transactions on Biomedical Engineering*, 45(3):277–286, March 1998.

3. M. Basseville. Distance measures for signal processing and pattern recognition. *Signal Processing*, 18(4):349–369, December 1989.
4. A. Bissacco, A. Chiuso, Y. Ma, and S. Soatto. Recognition of human gaits. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 01)*, volume II, pages 52–58, Kauai, Hawaii, December 2001.
5. J. Boets, K. De Cock, and B. De Moor. Distances between dynamical models for clustering time series. In *Proceedings of the 14th IFAC Symposium on System Identification (SYSID 2006)*, pages 392–397, Newcastle, Australia, March 2006.
6. J. Boets, K. De Cock, M. Espinoza, and B. De Moor. Clustering time series, subspace identification and cepstral distances. *Communications in Information and Systems*, 5(1):69–96, 2005.
7. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. Wiley, New York, 1991.
8. K. De Cock. *Principal Angles in System Theory, Information theory and Signal Processing*. PhD thesis, K.U.Leuven, Leuven, Belgium, May 2002. Available as "ftp://ftp.esat.kuleuven.be/pub/SISTA/decock/reports/phd.ps.gz".
9. K. De Cock and B. De Moor. Subspace angles between ARMA models. *Systems & Control Letters*, 46(4):265–270, July 2002.
10. U. B. Desai, D. Pal, and R. D. Kirkpatrick. A realization approach to stochastic model reduction. *International Journal of Control*, 42(4):821–838, 1985.
11. G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003.
12. W. Gersch. Nearest neighbor rule in classification of stationary and nonstationary time series. In D. F. Findley, editor, *Applied Time Series Analysis II*, pages 221–270. Academic Press, New York, 1981.
13. R. M. Gray, A. Buzo, A. H. Gray, Jr., and Y. Matsuyama. Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–28(4):367–376, August 1980.
14. A. H. Gray, Jr. and J. D. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–24(5):380–391, October 1976.
15. E. J. Hannan and D. S. Poskitt. Unit canonical correlations between future and past. *The Annals of Statistics*, 16(2):784–790, June 1988.
16. H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–372, 1936.
17. I. A. Ibragimov and Y. A. Rozanov. *Gaussian Random Processes*. Springer, New York, 1978.
18. F. Itakura and T. Umezaki. Distance measure for speech recognition based on the smoothed group delay spectrum. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP87)*, volume 3, pages 1257–1260, 1987.
19. N. P. Jewell and P. Bloomfield. Canonical correlations of past and future for time series: definitions and theory. *The Annals of Statistics*, 11(3):837–847, 1983.
20. N. P. Jewell, P. Bloomfield, and F. C. Bartmann. Canonical correlations of past and future for time series: bounds and computation. *The Annals of Statistics*, 11(3):848–855, 1983.
21. C. Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société Mathématique*, 3:103–174, 1875.
22. Y. Kakizawa, R. H. Shumway, and M. Taniguchi. Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93:328–340, 1998.
23. K. Kalpakis, D. Gada, and V. Puttagunta. Distance measures for effective clustering of ARIMA time-series. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM'01)*, pages 273–280, San Jose, CA, November-December 2001.
24. T. Katayama and G. Picci. Realization of stochastic systems with exogenous inputs and subspace identification methods. *Automatica*, 35(10):1635–1652, 1999.

25. D. Kazakos and P. Papantoni-Kazakos. Spectral distance measures between Gaussian processes. *IEEE Transactions on Automatic Control*, 25(5):950–959, 1980.
26. L. Li and Z. Xie. Model selection and order determination for time series by information between the past and the future. *Journal of time series analysis*, 17(1):65–84, 1996.
27. A. Lindquist and G. Picci. Canonical correlation analysis, approximate covariance extension, and identification of stationary time series. *Automatica*, 32(5):709–733, 1996.
28. R. J. Martin. A metric for ARMA processes. *IEEE Transactions on Signal Processing*, 48(4):1164–1170, April 2000.
29. M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden–Day, San Francisco, 1964. Originally published in Russian in 1960.
30. F. C. Schweppe. On the Bhattacharyya distance and the divergence between Gaussian processes. *Information and Control*, 11(4):373–395, 1967.
31. F. C. Schweppe. State space evaluation of the Bhattacharyya distance between two Gaussian processes. *Information and Control*, 11(3):352–372, 1967.
32. R. H. Shumway and A. N. Unger. Linear discriminant functions for stationary time series. *Journal of the American Statistical Association*, 69:948–956, December 1974.
33. P. Van Overschee and B. De Moor. Subspace algorithms for the stochastic identification problem. *Automatica*, 29:649–660, 1993.
34. P. Van Overschee and B. De Moor. *Subspace Identification for Linear Systems: Theory – Implementation – Applications*. Kluwer Academic Publishers, Boston, 1996.