GEOMETRIC METHODS FOR STATE SPACE IDENTIFICATION

Anders Lindquist[1] and Giorgio Picci[2]

[1] Optimization and System Theory, Dept. of Mathematics, Royal Institute of Technology, S-10084 Sweden
[2] Department of Electronics and Computer Science, University of Padua, via Gradenigo 6/A, 35131 Padua, Italy and LADSEB-CNR, Padua, Italy

1. Introduction

The scope of identification theory is to construct algorithms for automatic model building from observed data. In these lectures we shall only discuss the case where the data are collected in one irrepetible experiment and no preparation of the experiment is possible (i.e. we cannot choose the experimental conditions or the input function to the system at our will).

The observed variables, usually classified as "inputs" (u) and "outputs" (y), are measured at discrete instants of time t and collected in a string of data of finite duration T. These data are called a "time series" in the statistical literature. There is a preselected model class, say the class of finite-dimensional linear time-invariant systems of a given order and the problem is generally formulated as that of inferring a "best" mathematical model in the model class on the basis of the observed data. There may be a variety of different reasons to build models. Here we shall be chiefly interested in model building for the purpose of prediction and control. This means that the identified model should be useful for prediction or control of *future* i.e. not yet observed, data.

Essential features of the Identification Problem.

1. There are always many other variables besides the preselected "inputs" and "outputs" which influence the time evolution of the system and hence the joint dynamics of y and u during the experiment. These variables represent the unavoidable interaction of the system with its environment. For this reason, even in the presence of a true causal relation between inputs and outputs there always are some *unpredictable* fluctuations of the values taken by the measured output $y(t)$ which are not explainable in terms of past input (and/or output) history.
   We cannot (and do not want to) take into account too many of these variables explicitely in the model as some of them may be inaccessible to measurement and in any case this would lead to complicated models with too many variables. We need to work with models of small complexity and treat the unpredictable fluctuations in some simple "aggregate" manner.

2. Models (however accurate) are of course always mathematical idealizations of nature. No physical phenomenon, even if the experiments were conducted in an ideal interactions-free environment can be described *exactly* by a set of differential or difference equations and even more so if the equations are a priori restricted to be linear, finite-dimensional and time-invariant. So the observables, even in an ideal "disturbance-free" situation cannot be expected to obey *exactly* any linear time-invariant model.

If we accept the arguments above it is clear that one essential issue to be addressed for a realistic formulation of the problem is a satisfactory notion of non-rigid, i.e. *approximate*, mathematical modeling of the observed data. The meaning of the word "approximate" should here be understood in the sense that a model should be able to accept as legitimate, data sets (time series) which may possibly differ slightly from each another. Imposing rigid "exact" descriptions of the type $F(u, y) = 0$ to experimental data has been criticized since the early beginnings of experimental science. Particularly illuminating is Gauss' general philosophical discussion in [27] sect. III, p. 236.

More to the point, there has been a widespread belief in the early years of control theory that identification was merely a matter of describing (exactly) the measured data by linear convolution equations of the type

$$y(t) = \sum_{t_0}^{t} h(t-\tau)u(\tau) \tag{1.1}$$

or, equivalently, by matching exactly pointwise harmonic response data with linear transfer function models. Results have always been poor and extremely sensitive to the data. New incoming data tend to change the model drastically, which means that a model determined in this way has in fact very poor predictive capabilities. The reason is that data obey exactly rigid relations of this kind "with probability zero". If in addition the model class is restricted to be finite-dimensional, which of course is what is really necessary for control applications, imposing the integral equation model (1.1) on real data normally leads to disastrous results. This is by now very well-known and documented in the literature, see e.g. [65, 70, 35, 20]. The fact, expressed in the language of numerical analysis, is that fitting rigid models to data invariably leads to ill-conditioned problems.

Gauss idea of describing data by a *distribution function* is a prime example of thinking in terms of (non-rigid) approximate models[1]. Other alternatives are possible, say using model classes consisting of a rigid "exact" model as a "nominal" object, plus an uncertainty ball around it. In this case, besides a "nominal" model, the identification procedure is required to provide at least

[1] A vulgar belief attributes to Gauss the invention of least squares, which is historically wrong. In Gauss' work least squares come out as a solution method for optimally fitting a certain class of *density functions* to the observed data.

bounds on the magnitude of the relative "uncertainty region" around the nominal model. This type of modeling philosophy is motivated by use in H^∞ control applications. Here one should provide a mathematical description of how the "dynamic" uncertainty ball is distributed in the frequency domain, rather than, as more traditionally done, in the parameter space, about the "nominal identified model".

In addition to the above we need also to introduce a mathematical description of *the data*. The data at our disposal at some fixed time instant represent only partial evidence about the behaviour of the system; we do not know the future continuation of the input and output time series, yet all possible continuations of our data must carry information about the same physical phenomenon we are about to model, and hence the possible continuations of the data cannot be "totally random" and must be related to what we have observed so far. So, in order to discover models of systems, we have to work with models of uncertain signals.

Mathematical descriptions of *uncertain signals* can be quite diverse. Possible choices are *stochastic processes*, deterministic signals with uncertainty bounds, etc. The crucial difference between theories of model building relates to the *quantitave* method for modeling uncertain signals they use[2].

In these lectures we shall eventually take the "classical" route and model uncertainty with the apparatus of probability theory. In this framework identification is phrased as a problem of mathematical Statistics.

One could argue that the basic problem of identification is, much more than designing algorithms which fit models to observed data (the easy part), the quantification of *dynamic uncertainty bounds* or the description of the *dynamic errors* incurred when using the model with future data. Any sensible identification method should provide some mathematical description of how uncertainty is distributed in time or frequency about the nominal identified model. In this respect the stochastic approach offers a very nice solution. In this setup (at least in the linear wide-sense setting) model uncertainty turns out to be equivalent to *additive* random disturbances i.e. identifying model uncertainty is equivalent to identifying models for "partially observed" stochastic processes. We shall discuss this point further in the following.

1.1 Stationary signals and the Statistical Theory of Model building

Since identification for the purpose of prediction and control makes sense only if you can use the identified model to describe future data, i.e. different data than those employed for its calibration, at the roots of any data-based model building procedure there must be a formalization of the belief that

[2] For this reason we would probably not classify as identification "exact modeling" where the data are "certain" signals assumed to fit exactly some finite set of (linear) relations.

future data will continue to be generated by the same "underlying mechanism" that has produced the actual data.

This is a vague but basic assumption on the nature of the data, which are postulated to keep being "statistically the same" in the future. Besides being inherent in the very *purpose* of collecting data for model building this assumption offers the logical background for assessing the *quality* of the identified model, by *asymptotic analysis*, i.e. by comparing finite-sample results with the "best achievable" model which could theoretically be identified with data of infinite lenght. One could probably say that *Statistics* as a discipline, is founded on asymptotic analysis, and that the wide use of Statistics and of probabilistic methods in identification is mainly motivated by the large body of effective asymptotic tools which can be applied to assess some basic "quality" features of the estimated model.

Classical Statistics traditionally starts by postulating some "urn model" whereby the data are imagined as being "drawn" at random from some universe of possible values in a "random trial" where "nature" chooses according to some probability law the current "state" of the interactions and of the experimental conditions.

It has been argued that the abstract "urn model" of probability theory looks inadequate to deal with situations like the one we have envisaged, where there is just one irrepetible experiment and there is really no sample space around from which the results of the experiment could possibly have been drawn. This critique comes from a tendendency to confuse physical reality with mathematical modelling. In effect the "urn model" is just a mathematical device which is not required to have any physical meaning or interpretation and can be used to model anything.

The critique has at least the merit of bringing up an important issue. It should be admitted that in large sectors of the literature the stochastic framework is often imposed dogmatically to practical problems (the user is normally left alone wondering if his problem is "stochastic" enough to be authorized to apply algorithm A, or his data are instead "determinstic" and he should apply algorithm B instead) and often statistical procedures are pushed to extremes where there really seems to be no physical ground for their applicability.

Yet there is a vast number of situations where a formal justification for the adoption of the probabilistic description of uncertain data can be given. Formal arguments leading to a probabilistic description of certain types of data could for example be based on the notion of "stationarity", a mathematical condition meant to capture the idea that future data should be "statistically the same" as past data. One possible line of reasoning is briefly elaborated upon below.

Let $z := \{z(t)\}_{t\in\mathbb{Z}}$ be a discrete-time signal (i.e. a sequence of real numbers). A *function of* z is just a real-valued function $f(z) := f(z(t); t \in I)$, $f : \mathbb{R}^I \to \mathbb{R}$ where I is a subinterval of $\mathbb{Z}$, possibly infinite. The *shift* σ is

the map defined on real sequences as $[\sigma z](t) := z(t+1), \quad t \in \mathbb{Z}$ so that the iterated application of σ, say

$$[\sigma^t z](s) := z(t+s), \quad t, s \in \mathbb{Z}$$

transforms a signal z into its "translation by t units of time" $z_t := \{z(t+s)\}_{s\in\mathbb{Z}}$. Let us denote by $f_t(z)$ the result of applying f to the shifted sequence $\sigma^t z$, i.e. let $f_t(z) := f(z(t+s); s \in I) = f(z_t), t \in\in \mathbb{Z}$.

Definition 1.1. *A signal z will be called*

– Strict-sense *stationary if the Cesaro limit*

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} f_t(z)$$

exists for all bounded measurable functions f;

– Wide-sense *stationary if the limit exists for $f(z) = z(0)$ (so that $f_t(z) = z(t)$) and for all quadratic forms*[3] *in z.*

The two conditions for wide-sense stationarity are normally found in the literature under a variety of different names. They describe the minimum amount of structure on the data which is necessary to do a (rudimental) asymptotic analysis of an identification algorithm for linear time-invariant models. The strict-sense notion is introduced mostly for conceptual reasons. Both notions generalize in a natural way to vector-valued sequences.

The purpose of the following paragraphs is to show that (strict-sense) stationary signals admit *stationary stochastic processes* as a natural mathematical description.

First take $f(z) := I_A(z(0))$ where I_A is the indicator function of a Borel set $A \subset \mathbb{R}$ ($I_A(x) = 1$ if $x \in A$ and 0 otherwise). Then the nonnegative number

$$\nu_T(A) := \frac{1}{T+1} \sum_{t=0}^{T} I_A(z(t))$$

is just the relative frequency of visits of the signal z to the set A. In fact, for each fixed T the function $A \to \nu_T(A)$ is a *probability measure*, i.e. a countably additive set function on the Borel sets of the real line. This follows simply from the relation $I_{\cup A_k} = \sum I_{A_k}$ which is valid for any sequence of disjoint sets A_k. For a stationary sequence we have $\nu_T(A) \to \nu_0(A)$ as $T \to \infty$. It then follows readily that

Lemma 1.1. *The set function $A \to \nu_0(A)$ is a probability measure on $\mathbb{R}$.*

[3] i.e. for all real functions f such that $f(\alpha z) = \alpha^2 f(z)$.

More generally, take

$$f(z) := I_A(z(0))I_{A_1}(z(\tau_1))\dots I_{A_n}(z(\tau_n))$$

where $\tau_1 \dots \tau_n$ are arbitrary time instants and $A, A_1 \dots A_n$ arbitrary Borel sets of the real line and consider the relative frequency

$$\nu_T(A, A_1, \tau_1, \dots A_n, \tau_n)) := \frac{1}{T+1}\sum_{t=0}^{T} I_A(z(t))I_{A_1}(z(t+\tau_1))\dots I_{A_n}(z(t+\tau_n))$$

of a visit to the set A followed by a visit, τ_1 instants later, to the set A_1, τ_2 instants later to the set A_2 etc.. and τ_n instants later to the set A_n. By stationarity $\nu_T(A, A_1, \tau_1, \dots A_n, \tau_n) \to \nu_n(A, A_1, \tau_1, \dots A_n, \tau_n)$ as $T \to \infty$. An easy generalization of Lemma 1.1 leads to the following statement.

Lemma 1.2. *The set function* $(A \times A_1 \dots \times A_n) \to \nu_n(A, A_1, \tau_1, \dots A_n, \tau_n)$ *is a probability measure on* $\mathbb{R}^{n+1}$ *for all time lags* $\tau_1 \dots \tau_n$. *In fact the family* $\{\nu_k\}_{k\in\mathbb{Z}_+}$ *is a* consistent family of probability distributions *in the sense of Kolmogorov, i.e.*

$$\nu_n(A, A_1, \tau_1, \dots, \mathbb{R}, \tau_n) = \nu_{n-1}(A, A_1, \tau_1, \dots A_{n-1}, \tau_{n-1})$$

for all Borel sets $A, A_1 \dots, A_{n-1}$ *and time lags* $\tau_1 \dots, \tau_n$.

It follows by a famous theorem of Kolmogorov that there is a bona - fide probability measure ν on the "sample space" $\mathbb{R}^{\mathbb{Z}}$ of all real sequences, which is the (unique) extension of the family of finite dimensional distributions $\{\nu_k\}_{k\in\mathbb{Z}_+}$ associated to a stationary signal z by the construction illustrated above. This measure is invariant with respect to the shift σ acting on the sequences of $\mathbb{R}^{\mathbb{Z}}$. In other words, tha pair $(\mathbb{R}^{\mathbb{Z}}, \nu)$ (with the natural family of measurable sets) defines a *stationary stochastic process* $\mathbf{z}$.

The moral of the story is that every stationary signal can be interpreted in a canonical way as a "representative" trajectory of a stationary process[4]. In other words,

Proposition 1.1. *For a stationary signal* z *there always exists an "urn model" i.e. a probability space* $\{\Omega, \mathcal{A}, \mu\}$ *and a stationary process* $\mathbf{z} := \{\, z(t, \omega) \,|\, t \in \mathbb{Z}, \omega \in \Omega \,\}$ *defined on it such that* z *is a representative trajectory of* $\mathbf{z}$, *i.e.*

$$z(t) = \mathbf{z}(t, \bar{\omega}) \quad t \in \mathbb{Z}$$

for some elementary event $\bar{\omega}$ *in the "good" set of probability one guaranteed by Birkhoff's theorem.*

[4] It is well known that *almost all* trajectories of a stationary process $\mathbf{z}$ are stationary signals in the sense of Definition 1.1. This is essentially the famous D.G. Birkhoff's *ergodic theorem*, see e.g. Doob [19], p. 465. A "representative" trajectory is just a trajectory belonging to the set of trajectories of ν-probability one where the Cesaro sums converge. Note that the process $\mathbf{z}$ need not be ergodic (i.e. "metrically transitive" according to the old terminology).

So we are authorized if we wish, to think legitimately of a stationary sequence of data as being "drawn" from a population according to a stationary probability law. We shall call this probability measure the *true (probability) law* of the data.

All of the above is of course mostly of "theoretical interest" and only serves the purpose of justifying the introduction of probabilistic and statistical language in identification. Very often in practice one can make verifiable statements only about the first and second order moments of the observed data and so in the following we shall normally work under the assumption of *wide sense stationarity* only. Moreover we shall assume throughout that the time averages of all signals have been subtracted off so all data will be assumed to be *zero mean* hereafter. Hence a wide-sense stationary signal (which we shall now assume m-dimensional) is just a sequence z for which the limit

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} z(t+\tau)z(t)' := \Lambda_0(\tau) \tag{1.2}$$

exists for all $\tau \in \mathbb{Z}$.

Proposition 1.2 (Wiener). *The function $\Lambda_0 := \tau \to \Lambda_0(\tau)$ is a bona-fide covariance function (i.e. a symmetric positive definite matrix function)*

Proof. The function Λ_0 is the discrete-time version of $\phi(x)$ in Wiener's Generalized Harmonic Analysis [73].

From this result, much in the same spirit of the strict-sense Proposition 1.1 stated above, one may draw the conclusion that a wide-sense stationary signal admits as a atural probabilistic model a *stationary wide-sense stochastic process*. Here, following [19] "wide-sense process" means the equivalence class of stochastic processes (defined say on the probability space $(\mathbb{R}^m)^{\mathbb{Z}}$) with zero mean and all having the same covariance function. In certain cases it may be appropriate to take as a representative of the equivalence class the unique *Gaussian* process with (zero mean and) given covariance function. Of course the additional strict-sense probabilistic structure provides only illusory extra information (besides second-order) unless the data provide actual evidence for the choice of Gaussian distributions.

A blanket assumption during the rest of these notes will be that the input-output data extend in the future do form a stationary[5] signal z; we shall call Λ_0 the *true covariance* of this signal.

Remarks. Note that for (wide-sense) stationary signals which decay to zero as $T \to \infty$ the true covariance function is identically zero. This is not paradoxical, as a signal of this kind may intuitively be regarded as a "transient" phenomenon settling eventually to a zero steady state.

[5] "Stationary" will mean wide-sense stationary hereafter.

The *spectral distribution function* of the signal is a monotonic Hermitian matrix function F_0 defined on the unit circle of the complex plane $\{\zeta = e^{j\omega}\}$ by the "Fourier-like" representation formula valid for any covariance function

$$\Lambda_0(\tau) = \int_{-\pi}^{\pi} e^{j\omega\tau} dF_0(e^{j\omega}) \tag{1.3}$$

(Herglotz Theorem). If the $\Lambda_0(\tau)$ form a summable sequence (so that $\Lambda_0(\tau) \to 0$ as $\tau \to \infty$) the spectral distribution function admits a density Φ_0,

$$F_0(e^{j\omega_2}) - F_0(e^{j\omega_1}) = \int_{\omega_1}^{\omega_2} \Phi_0(e^{j\lambda}) \frac{d\lambda}{2\pi}$$

In general when the covariance function does not decay to zero, for example when there are periodic components in z, the distribution function has jumps and the density function describes only the absolutely continuous part of F_0. *Persistently exciting signals* of order n are classical examples of periodic stationary signals whose distribution function is a staircase function with exactly n jumps.

The statistical approach to identification. As we have argued in this section a reasonable mathematical description of the measured data is to model it as a finite tract of a trajectory of a stationary (wide-sense) stochastic process. The identification problem is then naturally formulated as the problem of recovering the "true" law of the process i.e. its true covariance or spectral distribution function[6] from the measured data. This of course is just the prototypical problem of Statistics.

Naturally the family of all possible "true descriptions" is an exceedingly general infinite-dimensional object and to make the problem solvable one has to choose, perhaps on the basis of some available a priori information, a manageable subclass which should be describable in terms of a finite number of real parameters. In fact, we shall ask that the model class should be compatible with *finite dimensional* prediction and control schemes. Although we keep the meaning of the term rather vague at this stage, it is very well-known that "finite-dimensional" wide-sense stationary processes can only be linear combinations of quasi-periodic (i.e. sums of sinusoids with random amplitudes) and purely-non-deterministic processes with a *rational spectral density*. There is then very little choice for the model class. If we are interested in finite-complexity models of "truly random" (purely-non-deterministic) signals, then we must restrict to *rational spectral densities*.

[6] Of course, more generally, in a stric-sense formulation one tries to "infer" the true probability law of the underlying process.

1.2 Input-Output models

Very often in "input-output" experiments one is not interested in modeling the input signals and would like to concentrate just on recovering a (causal) relation between inputs and outputs.

In the present wide-sense stochastic setup the input-output relation has in general a linear structure of the type

$$y(t) = E[\, y(t) \,|\, u(s); s \leq t \,] + v(t) \tag{1.4}$$

where $E[\, y(t) \,|\, u(s); s \leq t \,]$ is the best (in the sense of minimum variance of the error) estimate of the output $y(t)$ based on the past of u up to time t. By Wiener filtering theory it known that this estimate is described by a causal and stable linear convolution operator with a rational transfer function $F(\zeta)$. The additive term $v(t)$ is the relative "estimation error", a stationary process with rational spectrum, uncorrelated with the past of u, which models precisely the uncertainty due to disturbances etc. superimposed to the input-based prediction, $F(\zeta)u(t)$, of $y(t)$.

The structure of the "input-output" model class which results from the assumptions of joint wide-sense stationarity and rational joint spectrum for the input and output processes is quite explicit indeed. Note that it comes out, as a formal consequence of the probabilistic setting used to describe our data. There is no arbitrariness or "user choice" at this stage, except of course for the choice of the order or the structure parameters of the transfer function. Note incidentally that identifying the model uncertainty in (1.4) means identifying a dynamic model for the additive noise process v.

A typical route which is commonly taken is to estimate the transfer function F and the noise model for v as if u was a deterministic sequence. Sometimes in the literature it is even "assumed" that u is a "deterministic" signal. This of course cannot be the real intention since it would lead to the rather absurd consequences that

$$\mathrm{E} \sum_{t,s} y(t)u(s) = \sum_{t,s} [\mathrm{E}y(t)]\, u(s) = 0$$

i.e. the input and output signals would be *completely uncorrelated*.

Estimation of the nominal input-output transfer function is generally to be understood as being "conditional on the past observed history of u". Although this may at a first sight look like a reasonable thing to do, it may lead to serious errors whenever hidden feedback links are present influencing the way in which the input variable is manufactured (i.e. introducing in u "stochastic components" correlated with the past of y).

In fact, if there is feedback from y to u the very notion of "input" looses its meaning since, as shown e.g. in [29], the input variable $u(t)$ is then also determined by a dynamical relation of the form (1.4), involving now the "output" process y playing in turn the role of an exogenous variable ("input") to determine u.

The appropriate setup for discussing these matters is within the theory of *feedback* and *causality* between stationary processess [33]. We shall not adventure into this subject in these introductory notes. We shall just content ourselves of recalling, as it has been argued in several places in the literature, that identification in the presence of feedback (and of course in the absence of any other specific information on the feedback loop) is essentially equivalent to identification of the *joint process* $[y', u']'$, in the sense of modeling the joint dynamics of the signals on the basis of the observed time-series $\{[y(t)', u(t)']'\}$. It is also for these reasons that we shall choose to restrict the scope of our discussion only to time-series identification in the rest of the paper.

2. State Space Models of Stationary processes

From the previous section it has been seen that wide-sense stationary processes with a *rational spectral density matrix* provide a natural class of finitely-parametrized stochastic models for the identification of a wide class of observed data.

It is very well known that these processes are precisely those admitting finite-dimensional *state-space descriptions* (or *realizations*) with constant parameters. It is then natural to pose the identification problem directly in terms of recovering state-space models of y. There are different approaches to identify models of this kind and we shall discuss some recent methods (the so-called "subspace methods") in some detail later in sections 6 and 7.

In any case it is well-known that even if we restrict to *minimal models* i.e. models of the smallest possible dimension of the state space, there are in general many non-equivalent (minimal) state-space representations of the same process y. This is a significant departure from the usual deterministic linear modeling setup and brings up model choice or *identifiability* questions which should be understood well before discussing the choice of a particular statistical methodology for model building. Therefore in this and in the following two sections we shall have to review the basic facts about finite-dimensional state-space models of stationary random processes.

Consider a stationary stochastic system

$$(\Sigma) \quad \begin{cases} x(t+1) &= Ax(t) + Bw(t) \\ y(t) &= Cx(t) + Dw(t) \end{cases} \tag{2.1}$$

where $\{w(t)\}$ is p-dimensional normalized white noise , i.e.

$$\mathrm{E}\{w(t)w(s)'\} = I\delta_{ts} \quad \mathrm{E}\{w(t)\} = 0.$$

In this paper we shall think of (2.1) exclusively as a *representation* of the output process y. This representation involves *auxiliary variables* such as the

state process x and the *generating white noise* w which are introduced for the purpose of giving the model a particular structure. These auxiliary variables play the role of "parameters" which may be eliminated producing a different model structure. For example, by eliminating x from the equations (2.1) one obtains an "input-output" representation whereby y appears as the result of processing the white noise signal w through a linear time-invariant filter

$$\xrightarrow{w} \boxed{\text{W}} \xrightarrow{y} \tag{2.2}$$

of transfer function

$$W(z) = C(zI - A)^{-1}B + D. \tag{2.3}$$

We shall for the moment make the assumption that the matrix A is *stable*, i.e. the eigenvalues of A all lie inside the unit circle ($|\lambda(A)| < 1$) and that the input noise has been applied to the system for an infinitely long time, i.e. starting at $t = -\infty$. In these conditions the effect of initial conditions has died off and the system is in statistical steady state. Then

$$x(t) = \sum_{j=-\infty}^{t-1} A^{t-1-j}Bw(j)$$

and

$$y(t) = \sum_{j=-\infty}^{t-1} CA^{t-1-j}Bw(j) + Dw(t)$$

In particular, x and y are *jointly stationary*[7].

The system 2.1 can be regarded as a linear map defining x and y as linear functionals of the input noise w. In fact, since the matrix A has been assumed stable, this map will be a *causal* map. In order to capture these properties in a precise way it is convenient to think of the (components of) x and y as elements of the infinite dimensional Hilbert space of second order random variables

$$H(w) = \overline{\text{span}}\{w_i(t) \mid t \in \mathbb{Z}; i = 1, 2, \ldots, p\} \tag{2.4}$$

with inner product $(\xi, \eta) = \text{E}\{\xi\eta\}$. Here $\overline{\text{span}}$ denotes the closure of the vector space generated by linear combinations of the elements listed inside the brackets. The Hilbert space $H(w)$ is called the *ambient space* of the stochastic system (Σ). It comes equipped with a unitary *Shift operator* U which is the extension of temporal translation i.e. $Ua'w(t) = a'w(t+1)$ of the generating random variables $a'w(t)$ of the space. More generally, the symbol $H(y)$ is used to denote the Hilbert space generated by a wide-sense zero mean process y. If the process is stationary then $H(y)$ is equipped with

[7] Stationarity here is always meant in the "wide sense" of second order statistics. In particular x and y being jointly stationary means that the covariance matrix $\text{E}\{[x(t)'y(t)']'[x(s)'y(s)']\}$ depends only on $t - s$.

the unitary shift of the process, U. The pair $(H(y), U)$ is called a *stationary Hilbert space*. By definition a stationary Hilbert space contains all translates $U^t\xi$ of any random variable ξ which belongs to it.

The *past subspaces* of x and y

$$H_t^-(x) = \overline{\text{span}}\{x_i(s) \mid s < t; i = 1, 2, \dots, n\} \tag{2.5}$$
$$H_t^-(y) = \overline{\text{span}}\{y_i(s) \mid s < t; i = 1, 2, \dots, m\} \tag{2.6}$$

are both contained in $H_t^-(w)$ (causality) and hence the *future space* of w

$$H_t^+(w) = \overline{\text{span}}\{w_i(s) \mid s \geq t; i = 1, 2, \dots, m\}$$

will be orthogonal to (i.e. uncorrelated with) both $H_t^-(x)$ and $H_t^-(y)$.

The finite dimensional subspace of $H(w)$

$$X_t = \text{span}\{x_1(t), x_2(t), \dots, x_n(t)\} \quad t \in \mathbb{Z},$$

is called the *state space* of the system 2.1 at the instant t.

In the following we shall always suppose that (A, B, C, D) in (2.3) is a minimal realization of W. In other words we shall assume that (A, B) is reachable and (C, A) is observable. Then, setting

$$P = \text{E}\{x(0)x(0)'\},$$

it follows from stationarity that $P = \text{E}\{x(t)x(t)'\}$ for all t, and hence the first equation in 2.1 yields

$$P = APA' + BB', \tag{2.7}$$

which is a Lyapunov equation. Since $|\lambda(A)| < 1$ the sum $P = \sum_{j=0}^{\infty} A^j BB'(A')^j$, converges and converges to the reachability grammian of Σ. But (A, B) is reachable, and hence $P > 0$. This implies that $\{x_1(t), x_2(t), \dots, x_n(t)\}$ is a *basis* in X_t.

Notations. We shall use the symbol $\vee$ to denote vector sum of subspaces, $+$ to denote *direct sum* and $\oplus$ to denote *orthogonal* vector sum. The orthogonal complement of a subspace A in the ambient space under consideration will be denoted by $A^\perp$. The future spaces always contain the present while the past does not [this convention will be followed generally with the only exception of Markov processes where both past and future must contain the present].

Several subspace constructions in the following are defined at some fixed reference time; by stationarity however they carry over to arbitrary time instants and we shall always implicitly mean that the relevant definition is extended by stationarity to the whole time axis.

Normally the reference time will be taken to be $t = 0$. To simplify notations the subscript $t = 0$ will normally be dropped. The symbols H^+ and H^- will denote the future and past spaces at time 0 of the process y. The orthogonal projection onto a subspace S will be denoted E^S or $\text{E}[.|S]$. For Gaussian random variables this coincides with the conditional expectation given the σ-algebra generated by S. Operators like E^S or U applied to vectors will act componentwise in an obvious way.

The Coordinate-free viewpoint. The coordinate-free or *geometric* viewpoint lies at the grounds of the identification methods which will be discussed in the last sections.

The main idea here is that building state-space models of a random process (i.e. stochastic realization) is essentially a matter of constructing a space X with properties which make it the stochastic analog of a deterministic state space. Once this first basic step is done, the rest is just a matter of choosing coordinates in X and the causality structure of the model. The basic notion in this respect is the following.

Definition 2.1. *Let X be a subspace of some large stationary Hilbert space H of wide-sense random variables containing $H(y)$. Define*

$$X_t := U^t X, \quad X_t^- := \vee_{s \le t} X_s, \quad X_t^+ := \vee_{s \ge t} X_s.$$

A Markovian Splitting Subspace *X for the process y is a subspace of H making the vector sums $H^- \vee X^-$ and $H^+ \vee X^+$ conditionally orthogonal (i.e. uncorrelated) given X, denoted,*

$$H^- \vee X^- \perp H^+ \vee X^+ \,|\, X. \tag{2.8}$$

The subspace X is called proper*, or* purely-non-deterministic *if there are vector white noise processes w and $\bar{w}$ such that* [8]

$$H^- \vee X^- = H^-(w), \quad H^+ \vee X^+ = H^+(\bar{w})$$

Any basis vector $x(0) := [x_1(0), x_2(0), \dots, x_n(0)]'$ in a Markovian splitting subspace X generates a stationary Markov process $x(t) := U^t x(0), t \in \mathbb{Z}$ which serves as a *state* of the the process y. If X is proper the Markov process is purely non determinstic and can be represented by a linear equation of the type $x(t+1) = Ax(t) + Bw(t)$ where A has all its eigenvalues strictly inside of the unit circle.

The fundamental characterization in this setting is the following.

Theorem 2.1. *[66, 47, 49] The state space X of any stochastic ralization (2.1) is a* Markovian Splitting Subspace *for the process y.*

Conversely, given any proper Markovian splitting subspace X, to any choice of basis $x(0) = [\, x_1(0), x_2(0), \dots, x_n(0)\,]'$ in X there corresponds a stochastic ralization of y of the type (2.1) with generating input noise w.

[8] This is equivalent to requiring that

$$\cap_t H_t^- \vee X_t^- = \{0\}, \text{ and } \cap_t H_t^+ \vee X_t^+ = \{0\}.$$

Obviously in this case y must also be purely non determinstic [68].

There are formulas expressing the coefficient matrices A, B, C, D in terms of x and y. They will be given in Theorem 2.2 below.

A Markovian splitting subspace is *minimal* if it doesn't contain (properly) other Markovian splitting subspaces. Countrary to the deterministic situation minimal Markovian splitting subspaces are *non unique*. Two very important examples are the *forward and backward predictor spaces* (at time zero):

$$X_- := \mathrm{E}^{H^-} H^+ \quad X_+ := \mathrm{E}^{H^+} H^- \tag{2.9}$$

for which we have the following characterization [49].

Proposition 2.1. *The subspaces X_- and X_+ are the unique minimal splitting subspaces contained in the past H^-, and, respectively, in the future H^+, of the process y.*

The causality of the representation (2.1) can be expressed geometrically as the orthogonality relation

$$H_t^+(w) \perp X_t^- \vee H_t^-(y) \tag{2.10}$$

for all $t \in \mathbb{Z}$. One also says that Σ is a *forward* model or that it evolves *forward* in time. Note in particular, that $\mathrm{E}\{x(t)w(t)'\} = 0$.

Backward or Anticausal realizations are models where instead the past of the driving white noise is orthogonal to the future of the state and output processes. These models are useful in several instances and are as legitimate representations of y as the forward models studied so far. As a matter of fact, a random signal has no "preferred direction of time" or causality built in and admits many different sorts of causality structures, see [67].

Theorem 2.2. *[48, 47] Any choice of basis vector $x(0)$ in a (finite dimensional) proper Markovian splitting subspace X generates a stationary vector Markov process $x(t) = U^t x(0)\,, t \in \mathbb{Z}$ such that the joint process $\begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix}$ is also Markov. The joint process admits a* forward representation

$$\begin{bmatrix} x(t+1) \\ y(t) \end{bmatrix} = \begin{bmatrix} A & 0 \\ C & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ y(t-1) \end{bmatrix} + \begin{bmatrix} B \\ D \end{bmatrix} w(t) \tag{2.11}$$

where $w(t)$ is the generating white noise process of $H^- \vee X^-$ i.e. $H^- \vee X^- = H^-(w)$ and

$$A = Ex(t+1)x(t)'P^{-1} \qquad B = Ex(t+1)w(t)' \tag{2.12}$$

$$C = Ey(t)x(t)'P^{-1} \qquad D = Ey(t)w(t)' \tag{2.13}$$

Dually, let $\bar{x}(0)$ be another basis in X and let $\bar{x}(t) = U^t \bar{x}(0)\, t \in \mathbb{Z}$ be the corresponding stationary vector Markov process. The joint process $\begin{bmatrix} \bar{x}(t) \\ y(t) \end{bmatrix}$ is also Markov and admits a backward representation

$$\begin{bmatrix} \bar{x}(t-1) \\ y(t-1) \end{bmatrix} = \begin{bmatrix} \bar{A} & 0 \\ \bar{C} & 0 \end{bmatrix} \begin{bmatrix} \bar{x}(t) \\ y(t) \end{bmatrix} + \begin{bmatrix} \bar{B} \\ \bar{D} \end{bmatrix} \bar{w}(t-1) \tag{2.14}$$

where $\bar{w}(t)$ is the generating white noise process of $H^+ \vee X^+$, i.e. $H^+ \vee X^+ = H^+(\bar{w})$ and

$$\bar{A} = E\bar{x}(t-1)\bar{x}(t)'\bar{P}^{-1} \qquad \bar{B} = E\bar{x}(t)\bar{w}(t)' \tag{2.15}$$

$$\bar{C} = Ey(t-1)\bar{x}(t)'\bar{P}^{-1} \qquad \bar{D} = Ey(t)\bar{w}(t)' \tag{2.16}$$

where $\bar{P} = E\bar{x}(t)\bar{x}(t)'$.

Taking $\bar{x}(t)$ as the dual basis *of $x(t)$, i.e.*

$$E\bar{x}(t)x(t)' = I$$

which implies

$$\bar{x}(t) = P^{-1}x(t), \quad \bar{P} = P^{-1},$$

the matrices of the backward representation $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ are related to (A, B, C, D) by a one-to-one transformation. In particular,

$$\bar{A} = A' \quad \bar{C}' = APC' + BD' \tag{2.17}$$

The formulas are asymmetric because of the asymmetry in the definition of past and future of y. This asymmetry is needed in order to avoid unnecessarily high state space dimension due to the overlap of past and future spaces of the process. For example, with the symmetric choice of including the present both in the future and in the past, a p-dimensional white noise process w (which is Markov) would admit its present space (spanned by $w(0)$) as a minimal Markovian splitting subspace and hence admit a minimal realization with a state space of dimension p. The choice here is to have the present only in $H^+(y)$.

3. Spectral Factorization

The covariance sequence of the output process y of the system (2.1), i.e.

$$\Lambda(t) := \mathrm{E}\{y(t+k)y(k)'\} = \mathrm{E}\{y(t)y(0)'\}$$

is readily computed. We see that

$$\Lambda(t) = CA^{t-1}\bar{C}' \quad \text{for } t > 0, \quad \Lambda(0) = CPC' + DD' \tag{3.1}$$

where,

$$\bar{C}' = APC' + BD'. \tag{3.2}$$

is exactly the same "backward" C matrix of (2.17) and

$$\Lambda(-t) = \Lambda(t)' = \bar{C}(A')^{t-1}C' \quad \text{for } t > 0.$$

Therefore it follows that the infinite block Hankel matrix

$$\mathbb{H} := \begin{bmatrix} \Lambda(1) & \Lambda(2) & \Lambda(3) & \cdots \\ \Lambda(2) & \Lambda(3) & \Lambda(4) & \cdots \\ \Lambda(3) & \Lambda(4) & \Lambda(5) & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

admits a facorization

$$\mathbb{H} = \begin{bmatrix} C \\ CA \\ CA^2 \\ CA^3 \\ \vdots \end{bmatrix} \begin{bmatrix} \bar{C} \\ \bar{C}A' \\ \bar{C}(A')^2 \\ \bar{C}(A')^3 \\ \vdots \end{bmatrix}' \tag{3.3}$$

and hence has finite rank bounded above by the dimension n of the state space X_t of the system Σ. Whether or not $\operatorname{rank}\mathbb{H} = n$ depends on the reachability of the pair $(A, \bar{C}')$, which is equivalent (given that (A, C) is observable by assumption) to *stochastic minimality* of the system (2.1) viewed as a representation of the output process y [47, 49, 50]. Note that

Proposition 3.1. *The backward state-output matrix $\bar{C}$ is uniquely determined by the forward parameters (A, C) and is invariant for all stochastically minimal realization (2.1) of y having the same (observable) (A, C) pair.*

We noted in the previous section that the output process y of (2.1) is a purely non-deterministic process. It is well known that this property is equivalent to

$$a'y(t) \notin H_t^-(y) \quad a \in \mathbb{R}^n.$$

i.e. for no $a \in \mathbb{R}^n$ $a'y(t)$ can be exactly equal to a linear combination of components of past variabes $y(t-1), y(t-2), \ldots$ of the process. From this it can be easily shown that the block Toeplitz matrix

$$T_k := \begin{bmatrix} \Lambda(0) & \Lambda(1) & \Lambda(2) & \cdots & \Lambda(k) \\ \Lambda(1)' & \Lambda(0) & \Lambda(1) & \cdots & \Lambda(k-1) \\ \Lambda(2)' & \Lambda(1)' & \Lambda(0) & \cdots & \Lambda(k-2) \\ \vdots & \vdots & \vdots & & \\ \Lambda(k)' & \Lambda(k-1)' & \Lambda(k-2)' & \cdots & \Lambda(0) \end{bmatrix} \tag{3.4}$$

is *(strictly) positive definite for all k.*

For a purely non-deterministic process the spectral distribution is absolutely continuous [68] and admits a density. In our case the $m \times m$ *spectral density* of y can even be computed as an ordinary Fourier transform i.e.

$$\Phi(z) = \sum_{t=-\infty}^{\infty} \Lambda(t) z^{-t}.$$

Since A is stable the series is absolutely convergent in a neighborhood of the unit circle $\{|z| = 1\}$ of the complex plane and clearly has the property

$$\Phi(1/z) = \Phi(z)'$$

which sometimes is called *para-Hermitian symmetry*. We may write

$$\Phi(z) = \Phi_+(z) + \Phi_+(1/z)' \tag{3.5}$$

where $\Phi_+(z)$ is the "causal" (i.e. analytic outside of the unit circle) component of $\Phi(z)$, given by

$$\begin{aligned} \Phi_+(z) &= \frac{1}{2}\Lambda(0) + \Lambda(1)z^{-1} + \Lambda(2)z^{-2} + \cdots & (3.6)\\ &= C(zI - A)^{-1}\bar{C}' + \frac{1}{2}\Lambda(0). & (3.7) \end{aligned}$$

The positivity condition of the sequence of Toplitz matrices (3.4) is equivalent to positive semidefiniteness of $\Phi(z)$ on the unit circle i.e.

$$\Phi_+(e^{j\theta}) + \Phi_+(e^{-j\theta})' \geq 0 \quad \theta \in [-\pi, \pi] \tag{3.8}$$

which can be rewritten as $\Re e\Phi_+(e^{j\theta}) \geq 0$. From this, since Φ_+ has by construction all of its poles strictly inside the unit circle it is seeen that it is a *positive real* function. We shall call Φ_+ the *positive real part* of Φ.

Proposition 3.2. *The transfer function W of any state space representation of the process y of the type (2.1) is a* spectral factor *of Φ, i.e.*

$$W(z)W(1/z)' = \Phi(z). \tag{3.9}$$

There is a very straightforward proof of this result in case of a stable A matrix, based on the well-known formula for computing the output spectrum of a linear time-invariant filter with stationary input (this formula is sometimes called the Wiener-Kintchine theorem).

There is however also a purely algebraic proof based on an astute decomposition of the product $W(z)W(1/z)'$ which works in general for proper rational transfer functions and does not require stability of A and stationarity of the signals involved (of course in this case the "spectrum" $\Phi(z)$ is *defined* by the formulas (3.5) plus (3.7) and need not have a probabilistic meaning). The decomposition is based on a famous trick apparently invented by Kalman and Yakubovich, namely the identity

$$P - APA' = (zI - A)P(z^{-1}I - A') + (zI - A)PA' + AP(z^{-1}I - A'). \tag{3.10}$$

Proof. A straightforward calculation shows that

$$\begin{aligned} W(z)W(1/z)' &= [C(zI-A)^{-1}B+D][B'(z^{-1}I-A')^{-1}C'+D'] \\ &= C(zI-A)^{-1}BB'(z^{-1}I-A')^{-1}C' \\ &+ C(zI-A)^{-1}BD'+DB'(z^{-1}I-A)^{-1}C'+DD' \end{aligned}$$

so, in view of (2.7), (3.10),

$$\begin{aligned} W(z)W(1/z)' &= CPC'+DD'+C(zI-A)^{-1}(APC'+BD') \\ &+ (CPA'+DB')(z^{-1}I-A')^{-1}C' \\ &= \Phi_+(z)+\Phi_+(1/z)'. \end{aligned} \tag{3.11}$$

where the last equality follows from (3.2).

Note that (3.11) only requires existence of a solution to the Lyapunov equation $P = APA' + BB'$. In case A is stable this is of course guaranteed. In addition, W has all its poles inside the unit circle. Such a W is called a *stable* or, better, *analytic spectral factor*.

We shall need to consider also *antistable* or *coanalytic* (i.e. analytic in $\{|z| < 1\}$) spectral factors i.e. (rational) solutions of $\bar{W}(z)\bar{W}(1/z)' = \Phi(z)$, having all poles ouside of the unit circle. These spectral factors are in one-to-one correspondence with the stable factors $G(z)$ of the transpose spectrum $\Phi(z)'$ by the formula

$$\bar{W}(z) = G(1/z)$$

so that $\bar{W}(z)\bar{W}(1/z)' = G(1/z)G(z)' = \Phi(z)$.

By the same reasoning as done for stable spectral factors, antistable spectral factors turn out to be exactly the transfer functions of backward realizations of y, i.e. state-space representations of the form (2.14). For the transfer function of a backward model (2.14) can be written

$$\bar{W}(z) = \bar{C}(z^{-1}I-\bar{A})^{-1}\bar{B}+\bar{D}$$

where the $\bar{A}$ matrix is stable, i.e. has all eigenvalues inside of the unit circle. Since the realization (3.7) of Φ_+ induces a natural transpose realization for the transpose $\Phi_+(z)'$, namely

$$\Phi_+(z)' = \bar{C}(zI-A')^{-1}C'+\frac{1}{2}\Lambda(0), \tag{3.12}$$

we see that the dual choice of basis of Theorem 2.2 for the backward models is a natural one. Hence by just switching symbols according to the correspondence

$$A \leftrightarrow A' \quad C \leftrightarrow \bar{C},$$

one obtains characterizations of the family of antistable spectral factors and the corresponding backward models which are completely analogous to those for stable spectral factors and forward realizations.

An important observation to keep in mind is that even though we assumed minimality of the realization (A, B, C) in 2.1, the pair $(A, \bar{C}')$ *may not be reachable* and hence

$$\Phi_+(z) = C(zI - A)^{-1}\bar{C}' + \frac{1}{2}\Lambda(0)$$

may not be a minimal realization. This would imply that the McMillan degree of the spectrum is smaller than what appears from the spectral factorization equation (namely $2n$). In fact, we have the following proposition for the McMillan degrees of rational functions, whose proof can be found in Anderson's paper [5].

Proposition 3.3. *Let $\delta\{\cdot\}$ denote McMillan degree. Then:*

(i) If the rational matrices Z_1 and Z_2 have no poles in common, then

$$\delta(Z_1 + Z_2) = \delta(Z_1) + \delta(Z_2).$$

(ii) If W_1 and W_2 are rational matrix functions of compatible dimensions, then

$$\delta(W_1 W_2) \leq \delta(W_1) + \delta(W_2).$$

Applying this to

$$W(z)W(1/z)' = \Phi(z) = \Phi_+(z) + \Phi_+(1/z)',$$

we have

$$\delta(W) \geq \frac{1}{2}\delta(\Phi) = \delta(\Phi_+). \tag{3.13}$$

If we have equality, we say that W is a *minimal spectral factor* .

Well-known examples of minimal stable spectral factor are the *minimum phase* , sometimes also called the *outer*, and the *maximum phase* spectral factors, denoted $W_-(z)$ and $W_+(z)$ respectively. Both $W_-(z)$ and $W_+(z)$ are stable (i.e. analytic in $\{|z| \geq 1\}$) but the first has no zeros outside of the closed unit disk while the second has instead no zeros inside the open unit disk.

Dually, there are unique minimal antistable (or co-analytic) spectral factors with all the zeros outside or, respectively, inside of the unit circle, denoted[9] $\bar{W}_+$ and $\bar{W}_-$ respectively. The factor $\bar{W}_+$ is commonly called *conjugate minimum-phase* or *co-outer*.

Theorem 3.1. *All stable rational spectral factors can be constructed by post-multiplying the minimum phase factor by a stable rational matrix function $Q(z)$ such that*

$$Q(z)Q(z^{-1})' = I.$$

[9] The rationale for the subscripts will become clear from the *partial order* of realizations which we shall see in a moment.

Dually, all antistable rational spectral factors can be constructed by postmultiplying the minimum phase factor by an antistable rational matrix function $\bar{Q}(z)$ *such that*

$$\bar{Q}(z)\bar{Q}(z^{-1})' = I$$

Transfer function like Q or $\bar{Q}$ are called *all-pass* . Stable all-pass functions are called *inner*. The result above goes back to Youla's classical 1961 paper [77].

4. Spectral Factorization and the LMI

Let us now consider the following inverse problem: Given a proper rational spectral density Φ i.e. an $m \times m$ parahermitian matrix of (generic) full rank m, positive semidefinite on the unit circle, consider the problem of finding all *minimal stable* spectral factors W and the corresponding (minimal) realizations $W(z) = D + H(zI - F)^{-1}B$. (The condition that Φ is proper implies that all rational spectral factors are proper so that they have representations of this form). To solve this problem, first make the decomposition

$$\Phi(z) = \Phi_+(z) + \Phi_+(1/z)'$$

where $\Phi_+(z)$ has all its poles strictly inside the unit disk (so it is the positive real part of $\Phi(z)$) and compute a minimal realization

$$\Phi_+(z) = C(zI - A)^{-1}\bar{C}' + J,$$

where clearly

$$J + J' = \Lambda(0).$$

Note that A is a stable matrix. We shall solve the spectral factorization equation (3.9), giving a procedure to compute (F, H, B, D) from the "data" $(A, C, \bar{C}, \Lambda(0))$.

The problem can actually be reduced to finding just the B's and D's since F and H can be chosen equal for all factors.

Theorem 4.1. *Let* $(A, C, \bar{C}')$ *be a minimal realization. There is a one-to-one correpondence between minimal stable spectral factors of* $\Phi(z)$*, and symmetric* $n \times n$ *matrices* P *solving the* Linear Matrix Inequality

$$M(P) := \begin{bmatrix} P - APA' & \bar{C}' - APC' \\ \bar{C} - CPA' & \Lambda(0) - CPC' \end{bmatrix} \geq 0 \tag{4.1}$$

in the following sense:

Corresponding to each solution $P = P'$ *of (4.1), necessarily positive definite, consider the unique (modulo orthogonal transformations) full column rank matrix factor* $\begin{bmatrix} B \\ D \end{bmatrix}$ *of* $M(P)$*,*

$$M(P) = \begin{bmatrix} B \\ D \end{bmatrix} [B'D'] \tag{4.2}$$

and the rational matrix W parametrized in the form

$$W(z) = C(zI - A)^{-1}B + D. \tag{4.3}$$

Then (4.3) is a minimal realization of a stable minimal spectral factor of $\Phi(z)$. Conversely, for each stable minimal spectral factor W, with minimal realization $D + H(zI - F)^{-1}B$ we can choose a basis such that $F = A$ and $H = C$, and the corresponding pair $\begin{bmatrix} B \\ D \end{bmatrix}$ together with the solution $P = P'$ of the Lyapunov equation (2.7) satisfy the matrix equation (4.2) and hence the Linear Matrix Inequality (4.1).

Proof. Let $P = P'$ be a solution of (4.1) and B, D be computed as in (4.2). Then P solves the Lyapunov equation (2.7) and hence $P > 0$. Then forming the product $W(z)W(1/z)'$ it follows from the equation (3.11) above that $W = D + (zI - A)^{-1}B$ is a stable spectral factor. Note that (A, B) must be reachable for otherwise the McMillan degree of W, would be $\delta(W) < n = \frac{1}{2}\delta(\Phi)$ which contradicts (3.13). Therefore $W = D+(zI-A)^{-1}B$ is a minimal spectral factor.

To show the converse, assume $W = D+H(zI-F)^{-1}B$ is a minimal stable spectral factor. Then a $P = P' > 0$ exists solving the Lyapunov equation $BB' = P - FPF'$ and hence from the spectral factorization equation and the Kalman-Yakubovich identity we get

$$\begin{aligned}
&\Phi_+(z) + \Phi_+(1/z)' = W(z)W(1/z)' = \\
&\begin{bmatrix} H(zI-F)^{-1} & I \end{bmatrix} \begin{bmatrix} BB' & BD' \\ DB' & DD' \end{bmatrix} \begin{bmatrix} (z^{-1}I - F')^{-1}H' \\ I \end{bmatrix} = \\
&HPH' + DD' + H(zI-F)^{-1}(FPH' + BD') + \\
&+(HPF' + DB')(z^{-1}I - F')^{-1}H'.
\end{aligned}$$

We first argue that without loss of generality we can take $F = A$ and $H = C$. This follows readily since the equality above implies that Φ_+ is also realized by a matrix triple of the form (F, G, H). Now since we are considering only spectral factors for which $\delta(W) = \delta(\Phi_+)$, this realization must also be minimal and then the pairs (F, H) and (A, C) must be similar. In fact we may take $F = A$ and $H = C$ for all minimal spectral factors.

It is then obvious that (P, B, D) satisfy (4.2) and hence (4.1) has a positive definite solution (namely P).

The equations (4.2) are sometimes called the *positive real equations*, for reasons to be explained below, and can be written

$$\underbrace{\begin{bmatrix} P - APA' & \bar{C}' - APC' \\ \bar{C} - CPA' & \Lambda(0) - CPC' \end{bmatrix}}_{M(P)} = \begin{bmatrix} B \\ D \end{bmatrix} \begin{bmatrix} B' & D' \end{bmatrix} \geq 0$$

The linear function $M : \mathbb{R}^{n\times n} \to \mathbb{R}^{2n\times 2n}$ depends only on $(A, \bar{C}, C, \Lambda(0))$ which are given.

One immediate consequence of Theorem 4.1 is that the *dimension* of the minimal spectral factors can be computed from the rank of the corresponding matrix $M(P)$. In fact if we agree to keep

$$\operatorname{rank} \begin{bmatrix} B \\ D \end{bmatrix}$$

full, it follows immediately from the the factorization above that the corresponding $W(z)$ is $m \times p$ where $p = \operatorname{rank} M(P)$.

It can be shown [25] that the set of solutions to the LMI (4.1)

$$\mathcal{P} := \{P \mid P' = P,\, M(P) \geq 0\}$$

is closed, bounded and convex. Later we shall show that there are two special elements $P_-, P_+ \in \mathcal{P}$ so that

$$P_- \leq P \leq P_+ \quad \text{for all } P \in \mathcal{P}$$

where $P_1 \leq P_2$ means that $P_2 - P_1 \geq 0$ is positive semidefinite.

For completeness, we also state the following well-known result. We have made it appear as a corollary to Theorem 4.1 although historically things went quite the other way.

Positive Real Lemma (Kalman-Yakubovich-Popov). *The family $\mathcal{P}$ is nonempty if and only if Φ_+ is positive real, i.e. (3.8) holds.*

Therefore, in our case, $\mathcal{P} \neq \emptyset$.

The Dual Positive-Real Equations. A dual of Theorem 4.1 providing a one-to-one and onto pararmetrization of minimal antistable factors in terms of the solutions $\bar{P}$ of the *dual Linear Matrix Inequality*

$$\bar{M}(\bar{P}) := \begin{bmatrix} \bar{P} - A'\bar{P}A & C' - A'\bar{P}\bar{C}' \\ C - \bar{C}\bar{P}A & \Lambda(0) - \bar{C}\bar{P}\bar{C}' \end{bmatrix} \geq 0 \tag{4.4}$$

can readily be obtained by replacing the realization $\Phi_+(z) = C(zI-A)^{-1}\bar{C}' + J$, by the transpose realization representing $\Phi_+(z)'$ and repeating verbatim the proof above, see also [48].

Then to each $\bar{P} \in \bar{\mathcal{P}}$, solution set of the dual Linear Matrix Inequality (4.4) there corresponds an antistable minimal spectral factor

$$\bar{W}(z) = \bar{C}(z^{-1}I - A')^{-1}\bar{B} + \bar{D},$$

where $\bar{B}, \bar{D}$ are determined by the analog of the matrix factorization (4.2).

In the following we shall assume that

$$R(P) := \Lambda(0) - CPC' > 0 \tag{4.5}$$

for all $P \in \mathcal{P}$. This means that all minimal state space models of y have a full-rank additive noise term in the output equation ($DD' > 0$). This condition serves here only the purpose of avoiding the use of pseudo-inverses and of simplifying the exposition. We shall see in a moment that (4.5) implies that $\Phi(z)$ is (generically) of full rank m. It is curious that a natural characterization of the spectra for which this condition holds seems still to be an open question in the literature (see however the forthcoming paper [52]). Under this assumption, if $T := -(\bar{C}' - APC')R^{-1}$, a straight-forward calculation yields

$$\begin{bmatrix} I & T \\ 0 & I \end{bmatrix} M(P) \begin{bmatrix} I & 0 \\ T' & I \end{bmatrix} = \begin{bmatrix} -\Lambda(P) & 0 \\ 0 & R \end{bmatrix},$$

where

$$\Lambda(P) = APA' - P + (\bar{C}' - APC')R(P)^{-1}(\bar{C} - CPA), \tag{4.6}$$

Hence, $M(P) \geq 0$ if and only if P satisfies the *Riccati inequality*

$$\Lambda(P) \leq 0, \tag{4.7}$$

and

$$p = \operatorname{rank} M(P) = m + \operatorname{rank} \Lambda(P).$$

If P satisfies the algebraic Riccati equation

$$\Lambda(P) = 0, \tag{4.8}$$

$\operatorname{rank} M(P) = m$, the corresponding spectral factor W is *square* $m \times m$. These P form a subfamily $\mathcal{P}_0$ in $\mathcal{P}$. If $P \notin \mathcal{P}_0$, W is rectangular.

From spectral factors to stochastic realizations. We now examine the converse of Proposition 3.2. Let W and $\bar{W}$ be two minimal square stable and antistable spectral factors. It is easy to see that such factors play the role of transfer functions of "shaping filters" of the type (2.2) for the process y. To see this we just need to manifacture two white noise processes w and $\bar{w}$ serving as input white noise processes in the two filters, the filter with transfer function W being causal (stable) and the other anticausal and hence represented in the time domain by a convolution operator integrating the input "backwards in time". Since W and $\bar{W}$ are square and invertible transfer functions, the white noise processes can be generated by passing y through the "whitening filters" W^{-1} and $\bar{W}^{-1}$. (The general idea is just the same as the classical "whitening-shaping" filter dicotomy of Bode and Shannon.) It can be shown[10] that the whitened processes $w, \bar{w}$ are in fact well-defined functionals of the history of y.

In particular, since W_- is outer, the corresponding white noise process w_- is a causal functional of y, i.e. $w_-(t-1) \in H_t^-(y)$ for all t, so that we actually have $H_t^-(w_-) = H_t^-(y)$. For this reason, w_- is called the (normalized)

[10] A precise statement of this requires spectral representation theory and can not be given here. For a similar argument see Rozanov's book [68], chapter 7.

forward innovation process of y [74]. Similarly the white noise process $\bar{w}_+$ is an anticausal functional of y, i.e. $\bar{w}_+(t) \in H_t^+(y)$, so that $H_t^+(\bar{w}_+) = H_t^+(y)$ for all t; $\bar{w}_+$ is called the (normalized) *backward innovation* process of y.

Once the white noise inputs are determined, it is rather obvious that a minimal realization (A, B, C, D) of W and (from the dual LMI) a dual minimal ralization $(A', \bar{B}, \bar{C}, \bar{D})$ of $\bar{W}$ will provide two minimal state-space stochastic realizations of the process y, the first one being causal and the second anticausal. The state processes of the two realizations will have as covariance matrices the unique solutions of the Lyapunov equations (2.7) and, respectively,

$$\bar{P} = A'\bar{P}A + \bar{B}\bar{B}'$$

In force of Theorem 4.1, each solution P of the Positive-real equations (4.1) in $\mathcal{P}_0$ will then automatically be interpretable as the state covariance matrix of the state-space realization of y corresponding to the deterministic realization (A, B, C, D) of W. Dually any solution $\bar{P}$ of the dual Positive-real equations (4.4) will be the state covariance of a backward state-space realiztion. In other words, the P (and $\bar{P}$) matrices solutions of he LMI (for the time being belonging to the subsets $\mathcal{P}_0$ and $\bar{\mathcal{P}}_0$) have the meaning of *state covariances* of minimal forward and backward realizations of y.

This picture however generalizes also to all minimal (nonsquare) spectral factors. The only difficulty in the generalization is the nonuniqueness of the white generating noises w and $\bar{w}$ associated to the spectral factors. The difficulty can be overcomed by selecting the input noises in a fixed ambient space, which is small enough to make the w's unique but also big enough to allow a solution w of the convolution equation $y = Ww$ for each minimal spectral factor W (and $\bar{W}$).

Proposition 4.1. *[50] There exists a fixed "universal" stationary Hilbert space $H \supset H(y)$ such that for each minimal spectral factor W the convolution equation $y = Ww$ has a unique (modulo orthogonal transformations) solution w with the property $H(w) \subset H$.*

Assume that the dimension of a minimal stochastic realization of y is n. Then the "universal" stationary Hilbert space H can be chosen equal to the orthogonal sum

$$H = H(y) \oplus H(z)$$

where z is a fixed n-dimensional normalized white noise process uncorrelated with y.

Hence, in order to construct all possible minimal shaping filter representations of y we need, besides the process y itself, n additional "exogenous" independent white noise generators. The filters for the actual generation of the noise processes are discussed in [50], sect. 5.2.

Once we know how to construct the w or $\bar{w}$ processes we can associate to each minimal spectral factor W ($\bar{W}$) a minimal stochastic realization of y by just picking a (minimal) realization (A, B, C, D) of W (resp. a minimal

realization $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ of $\bar{W}$). The state vector of the realization will be a causal or anticausal functional in the past or future space of the white noise processes w or $\bar{w}$. Clearly if we choose all w 's in a fixed universal Hilbert space H, the state spaces X of these realizations will all be subspaces of H.

Theorem 4.2. *Let $(A, C, \bar{C}, \Lambda(0))$ be a minimal realization of $\Phi_+(z)$. There are bijective maps between the following sets*

1. *$\mathcal{W}$: The minimal stable spectral factors W (defined modulo right multiplication by a constant orthogonal matrix)*
2. *$\mathcal{P}$:The solutions P of the Linear Matrix inequality (4.1)*
3. *$\mathfrak{X}$: The minimal Markovian splitting subspaces for y in a fixed universal Hilbert space H as described in Proposition 4.1.*

Proof. The one-to-one and onto correspondence between $\mathcal{W}$ and $\mathcal{P}$ has been described already in Theorem 4.1. The correspondence between $\mathfrak{X}$ and $\mathcal{P}$ is established through a particular choice of basis in the state space which will be described in the next section.

Of course we have a dual version of this result in which the antistable spectral factors $\bar{\mathcal{W}}$ and the solutions of the dual Linear Matrix Inequality $\bar{\mathcal{P}}$ replace $\mathcal{W}$ and $\mathcal{P}$.

4.1 Ordering, (A,C) pairs and uniform choice of basis in $\mathfrak{X}$

In this section we shall make explicit the correspondence between state covariance matrices P solutions of the LMI (4.1) and Markovian splitting subspaces in $\mathfrak{X}$. In fact we shall establish a correspondence between P's and stochastic state-space realizations of y. This correspondence is intimately related to the notion of a *uniform choice of basis* in the family of minimal splitting subspaces $\mathfrak{X}$, which will be defined below. This notion will be useful for understanding the geometric approach to identification and the idea of stochastic balancing.

Recall [50, sect. 6] that in the family of minimal Markovian splitting subspaces $\mathfrak{X}$ one can introduce a natural partial order (denoted $\prec$), defined in terms of the *cosine of the angle* that each X makes with the future space[11] H^+, a subspace X_2 being "greater" than X_1 if it is "closer" (i.e. it makes a smaller angle) with the future than X_1.

According to this definition the forward and backward predictor spaces, X_- and X_+, defined in (2.9) are naturally the smallest and largest element in the family $\mathfrak{X}$ with respect to the partial order.

[11] The (cosine of the) "angle" between subspaces is defined e.g. in [4, vol. I] p. 69. It is just the smallest *canonical correlation coefficient* of the two subspaces of random variables X and H^+. The "angle" is the largest *principal angle* between the two subspaces. For these notions in a finite-dimensional setting one may consult e.g. [32].

Consider a minimal causal realization (2.1). By minimality the components of the n–vector $x(0)$ must form a basis in the relative splitting subspace X. Now, we recall from [49], [50], that two families of bases, say $\{x(0)\}$ and $\{\,\bar{x}(0)\,\}$, for the family $\mathcal{X}$ (i.e. each vector $x(0)$ is a basis in one X and similarly for each $\bar{x}(0)$) are called *uniformly ordered*, (or, for short, *uniform*) respectively in the *forward* or in the *backward sense*, if whenever $X_1 \prec X_2$ and $x_i(0)$ are bases in X_i, $(i = 1, 2)$, there holds

$$E^{X_1} x_2(0) = x_1(0) \tag{4.9}$$

or, respectively,

$$E^{X_2} \bar{x}_1(0) = \bar{x}_2(0) \tag{4.10}$$

the vectors $\bar{x}_i(0)$ being bases for the subspaces X_i, $(i = 1, 2)$.

Uniformly ordered bases can be constructed easily. For example pick a basis $x_+(0)$ in the "largest" state space X_+, then it can be easily seen by using the splitting property of X that the family

$$x(0) := \mathrm{E}^X x_+(0), \quad X \in \mathcal{X}$$

is a forward-uniform basis [50]. It also follows from the definition that for all bases $x(0)$ in a forward-uniform family, and, respectively, for all $\bar{x}(0)$ in a backward-uniform family we have the *invariant projection* property

$$E^{X_-} x(0) = x_-(0) \tag{4.11}$$

$$E^{X_+} \bar{x}(0) = \bar{x}_+(0) \tag{4.12}$$

where $x_-(0)$ is the basis relative to the forward predictor space X_- in the first family and $\bar{x}_+(0)$ is the basis relative to the backward predictor space X_+ in the second family.

Proposition 4.2. *[50] A forward-uniform choice of bases in $\mathcal{X}$ establishes a* lattice isomorphism *between $\mathcal{X}$ and the corresponding family of state covariance matrices $\mathcal{P} := \{P = \mathrm{E}x(0)x(0)' \,|\, x(0)$ basis in $X\}$, the latter set being endowed with the natural partial order of positive semidefinite matrices. This is equivalent to saying that $X_1 \prec X_2 \Leftrightarrow P_1 \leq P_2$.*

In a backward-uniform choice the ordering of the corresponding state covariance matrices $\bar{\mathcal{P}} := \{\bar{P} = \mathrm{E}\bar{x}(0)\bar{x}(0)' \,|\, \bar{x}(0)$ basis in $X\}$ is reversed, namely, $X_1 \prec X_2 \Leftrightarrow \bar{P}_2 \leq \bar{P}_1$.

Proof. We shall prove only the implication $\Rightarrow$. For the reverse consult [50], p. 282-283. Everything descends from the orthogonality

$$x_2(0) - \mathrm{E}^{X_1} x_2(0) = x_2(0) - x_1(0) \perp x_1(0)$$

which is equivalent to the defining (4.9). From this it follows that the covariance matrix of $x_2(0) - x_1(0)$ is equal to $P_2 - P_1$ and hence $P_2 \geq P_1$. The backward case follows by symmetry.

In particular, the bases $x_-(0)$ in X_- and $x_+(0)$ in X_+ in a forward-uniform choice, will have the smallest and, respectively, the largest state covariance matrices P_- and P_+ in $\mathcal{P}$. For a backward-uniform family it will instead happen that $\bar{P}_-$ is maximal and $\bar{P}_+$ is minimal.

That P_- and $\bar{P}_+$ are the covariance matrices of bases in the forward and backward predictor spaces can be seen also directly from the invariant projection property (4.11, 4.12). In fact $x_-(0)$ and $\bar{x}_+(0)$ are essentially the "forward and backward steady–state Kalman filter" estimates of any $x(0)$ (respectively, $\bar{x}(0)$) in a uniform family the bases. For, the first vector can be expressed as $x_-(0) := \mathrm{E}^{X^-}x(0)$ which is in turn equal to $\mathrm{E}^{H^-}x(0)$ by the splitting property. Hence $x(0) - x_-(0) \perp x_-(0) \in H^-$ so that $P - P_- \geq 0$ for all $P \in \mathcal{P}$. A completely analogous argument shows that $\bar{x}_+(0) = \mathrm{E}^{H^+}\bar{x}(0)$ and that $\bar{P} - \bar{P}_+ \geq 0$ for all $\bar{P} \in \bar{\mathcal{P}}$.

The most useful properties of uniform bases are summarized below.

Proposition 4.3. *[50] If $\{x(0)\}$ is a forward-uniform family of bases in $\mathcal{X}$ then, for each $x(0)$, the corresponding* dual basis *$\bar{x}(0)$, uniquely defined in X by the condition $Ex(0)\bar{x}(0)' = I$, defines a backward-uniform family in $\mathcal{X}$.*

Proposition 4.4. *[50] All* causal *minimal realizations of y corresponding to a forward-uniform family of bases are described by the same (A, C) pair. Conversely, a choice of bases in the subspaces $X \in \mathcal{X}$ yielding state equations with the same (A, C) pairs is forward-uniform.*

Likewise, all anticausal *realizations corresponding to a backward-uniform family of bases are described by the same $(\bar{A}, \bar{C})$ parameters and conversely.*

Therefore choosing the realizations of spectral factors in $\mathcal{W}$ in such a way that the A and C matrices are the same (Theorem 4.1) is exactly the same thing as choosing a forward uniform basis in $\mathcal{X}$. Dually, keeping the $\bar{A}, \bar{C}$ matrices invariant over the family of all minimal realizations of spectral factors $\bar{W} \in \bar{\mathcal{W}}$ is the same thing as having the corresponding state vectors $\bar{x}(0)$ related as in a backward uniform choice of basis.

The only backward-uniform family of bases we shall encounter in the following will be the *dual bases of a forward family.* Since these are related by the transformation $\bar{x}(0) = P^{-1}x(0)$ so that

$$\bar{P} = \mathrm{E}\bar{x}(0)\bar{x}(0)' = P^{-1} \tag{4.13}$$

it follows immediately that

Proposition 4.5. *The solution sets $\mathcal{P}, \bar{\mathcal{P}}$ of the Linear Matrix Inequality and of its dual (4.4) are related by the transformation $\bar{P} = P^{-1}$. In particular, the maximal element in $\mathcal{P}$ is $P_+ = \bar{P}_+^{-1}$.*

The following result says that choosing a uniform basis is after all as easy as choosing one basis in a particular X.

Proposition 4.6. *An arbitrary basis $x(0)$ in a minimal splitting subspace X can be uniquely extended to the whole family $\mathcal{X}$ in a uniform way (either in the forward or backward sense).*

Proof. First, compute $\bar{x}_+(0)$ by projecting $\bar{x}(0) := P^{-1}x(0)$ onto X_+, i.e., $\bar{x}_+(0) := \mathrm{E}^{X_+}P^{-1}x(0)$ and then go back to the "primal", i.e. let $x_+(0) = P_+^{-1}\bar{x}_+(0)$. Note that the projection of $x(0)$ onto X_+ is *not invariant* in a forward uniform basis.

Once $x_+(0)$ and (by a dual argument) $\bar{x}_-(0)$ are found, they can be used to generate two (dual families of uniform bases in $\mathcal{X}$, say $z(0)$ and $\bar{z}(0)$, by setting $z(0) := \mathrm{E}^X x_+(0)$, and $\bar{z}(0) := \mathrm{E}^X \bar{x}_-(0)$. It is immediate to check that $z(0)$ and $\bar{z}(0)$ are indeed dual bases and are related by the transformation $\bar{z}(0) = P^{-1}z(0)$ where $P = \mathrm{E}z(0)z(0)'$.

5. Finite-Interval realizations of a stationary process

Since in identification data are always finite we need to examine the problem of modeling the process y (which will still be assumed stationary and with a rational spectral density as in the previous two sections) on a *finite interval*. There seems to be no reason to worry about this since a stationary realization of the type (2.1) does of course describe the process also on any finite interval $[0, T]$. In this case however we must specify the *initial conditions vector* $x(0)$ at time zero, which is an essential extra parameter for a complete description of the process. The initial condition must in principle be estimated if the model class is chosen to consist of stationary realizations (2.1).

As we shall see there are *non-stationary* finite-interval realizations where the initial state is automatically fixed to zero and we do not need to worry about its estimation.

5.1 Forward and backward Kalman filtering and the family of minimal stationary realizations of y

Let us recall the structure of the Kalman filter for a minimal (forward) stochastic realization (2.1). Assume we are observing the output of the system starting at some finite time t_0 and define

$$H_{[t_0,t)} = \mathrm{span}\{a'y(s); a \in \mathbb{R}^m, t_0 \le s < t\}.$$

Then, the Kalman estimate

$$\hat{x}(t) = \mathrm{E}^{H_{[t_0,t)}}x(t) \tag{5.1}$$

is given by[12]

[12] Because of joint stationarity of x, y all time-varying quantities below depend on the time difference $t - t_0$. To keep notations simple we write this difference simply as t.

$$\hat{x}(t+1) = A\hat{x}(t) + K(t)[y(t) - C\hat{x}(t)]; \quad \hat{x}(t_0) = 0, \tag{5.2}$$

where the *Kalman gain matrix* $K(t)$ is given by

$$K(t) = [AQ(t)C' + BD'][CQ(t)C' + DD']^{-1}$$

$Q(t)$ being the error covariance matrix

$$Q(t) = \mathrm{E}\{[x(t) - \hat{x}(t)][x(t) - \hat{x}(t)]'\}$$

which is the solution of the *matrix Riccati equation*

$$\begin{array}{rcl} Q(t+1) & = & AQ(t)A' - \\ & - & [AQ(t)C' + BD'][CQ(t)C' + DD']^{-1}[AQ(t)C' + BD']' + BB' \\ Q(0) & = & P = \mathrm{E}\{x(0)x(0)'\} \end{array} \tag{5.3}$$

This Riccati equation depends on P, B and D, and consequently it varies with different realizations Σ, i.e. with different $P \in \mathcal{P}$. We shall now replace this Riccati equation with an invariant version, which depends only on the invariant parameters $A, C, \bar{C}$ and $\Lambda(0)$ of the spectral density of y. In this way we shall prove that the estimate (5.1) *is the same for all forward stationary realizations of* y.

To this end, note that from the orthogonality $x(t) - \hat{x}(t) \perp \hat{x}(t)$ it follows that $Q(t) = P - \Pi(t)$, where

$$\Pi(t) := \mathrm{E}\{\hat{x}(t)\hat{x}(t)'\}.$$

Then the Riccati equation (5.3) can be written

$$\begin{array}{rcl} P - \Pi(t+1) & = & APA' - A\Pi(t)A' \\ & - & [\bar{C}' - A\Pi(t)C'][\Lambda(0) - C\Pi(t)C']^{-1}[\bar{C}' - A\Pi(t)C']' + BB'. \end{array}$$

because $\bar{C}' = APC' + BD'$ and $\Lambda(0) = CPC' + DD'$. But P satisfies the Lyapunov equation

$$P = APA' + BB',$$

and therefore

$$\Pi(t+1) = \Pi(t) + \Lambda(\Pi(t)) \quad \Pi(0) = 0, \tag{5.4}$$

where $\Lambda : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is given by

$$\Lambda(P) = A'PA - P + (\bar{C}' - APC')(\Lambda(0) - CPC')^{-1}(\bar{C}' - APC')'.$$

This is precisely the same quadratic function Λ introduced in the previous section. Moreover we can write,

$$K(t) = [\bar{C}' - A\Pi(t)C'][\Lambda(0) - C\Pi(t)C']^{-1},$$

which also depends only on $(A, C, \bar{C}, \Lambda(0))$ and on the solution of the Riccati equation (5.4). The transient (forward) innovation process

$$e(t) := y(t) - C\hat{x}(t) \quad t \geq t_0,$$

is also invariant. It is in fact a white noise process of covariance matrix

$$R(t) = \Lambda(0) - C\Pi(t)C',$$

i.e. $\mathrm{E}\{e(t)e(s)'\} = R(t)\delta_{ts}$. Hence we have shown the following fact.

Proposition 5.1. *Let $(A, C, \bar{C}, \Lambda(0))$ be a minimial realization of the spectral density matrix of the stationary process y. Then y has a non-stationary realization*

$$\hat{x}(t+1) = A\hat{x}(t) + B(t)\epsilon(t) \quad \hat{x}(t_0) = 0, \tag{5.5}$$

$$y(t) = C\hat{x}(t) + D(t)\epsilon(t) \tag{5.6}$$

on $\{t \geq t_0\}$, where the state $\hat{x}(t)$ is the orthogonal projection onto $H_{[t_0,t)}$ (the minimum variance one-step ahead estimate) of the state $x(t)$ of any minimal stationary realization of y in the uniform basis induced by (A, C), $B(t) := K(t)R(t)^{1/2}$, $D(t) := R(t)^{1/2}$ and $\epsilon(t) := -R(t)^{-1/2}e(t)$ is a normalized m-dimensional white noise process: the normalized transient innovation process *of y on $\{t \geq t_0\}$.*

The state of the *Kalman Filter realization* (5.5, 5.6) is a *non-stationary* process. In effect, since $\mathrm{E}^{H_{[t_0,t)}} y(t+k) = \mathrm{E}^{H_{[t_0,t)}} CA^k x(t) = CA^k\hat{x}(t)$ for all $k \geq 0$, the components of $\hat{x}(t)$ span the "finite-memory" predictor space

$$\check{X}_{t-} := \mathrm{E}^{H_{[t_0,t)}} H_t^+. \tag{5.7}$$

Relation with the forward stationary innovation model. This is well-known. Since each realization Σ is a minimal realization, by standard Kalman Fitering theory $\Pi(t) \to \Pi_\infty > 0$ as $t \to \infty$. Then Π_∞ satisfies the algebraic Riccati equation

$$\Lambda(P) = 0$$

so since it is also symmetric and positive definite, $\Pi_\infty \in \mathcal{P}_0 \subset \mathcal{P}$. Moreover $Q(t) \to Q_\infty \geq 0$, where $Q_\infty = P - \Pi_\infty$, and hence

$$P \geq \Pi_\infty. \tag{5.8}$$

for all P and therefore we see that $P_- = \Pi_\infty$. Now let $t_0 \to -\infty$ in the Kalman filter. Then $\hat{x}(t)$ converges by the (wide-sense) martingale convergence theorem [18] to the *steady state Kalman filter* estimate $x_-(t) = \mathrm{E}^{H_t^-} x(t)$. Moreover both $B(t)$ and $D(t)$ converge to constant matrices which indeed satisfy the Positive-Real lemma equations for $P = P_-$.

Dually, we construct a *backward Kalman filter* based on a minimal backward model

$$\begin{array}{rcl} \bar{x}(t) & = & A'\bar{x}(t+1) + \bar{B}\bar{w}(t) \\ y(t) & = & \bar{C}\bar{x}(t+1) + \bar{D}\bar{w}(t) \end{array} \tag{5.9}$$

assuming we can observe $y(t)$ only in the interval $\{t \leq T\}$. Define the finite future space

$$H_{[t,T]} = \text{span}\{a'y(s); a \in \mathbb{R}^m, t \leq s \leq T\}.$$

and the backward Kalman estimate,

$$\hat{\bar{x}}(t) = \text{E}^{H_{[t,T]}}\bar{x}(t) \tag{5.10}$$

Then an analysis that is completely symmetric to the one presented above, projecting over the future, yields a backward Kalman filter which can be written as a backward non-stationary realization of the stationary process y,

$$\begin{array}{rcll} \hat{\bar{x}}(t) & = & A'\hat{\bar{x}}(t+1) + \bar{B}(t)\bar{\epsilon}(t) & \hat{\bar{x}}(T) = 0, \\ y(t) & = & \bar{C}\hat{\bar{x}}(t+1) + \bar{D}(t)\bar{\epsilon}(t) & \end{array} \tag{5.11}$$

on $\{t \leq T\}$, where $\bar{B}(t) := \bar{K}(t)\bar{R}(t)^{1/2}$, $\bar{D}(t) := \bar{R}(t)^{1/2}$ and $\bar{\epsilon}(t) := -\bar{R}(t)^{-1/2}\bar{e}(t)$ is a normalized m-dimensional white noise process: the *normalized backward transient innovation process* of y on $\{t \leq T\}$.

The *backward Kalman gain*,

$$\bar{K}(t) = [C' - A'\bar{\Pi}(t)\bar{C}'][\Lambda(0) - \bar{C}\bar{\Pi}(t)\bar{C}']^{-1}, \tag{5.12}$$

is now computed from the solution of the *dual Riccati equation*

$$\bar{\Pi}(t-1) = \bar{\Pi}(t) + \bar{\Lambda}(\bar{\Pi}(t)) \quad \bar{\Pi}(T) = 0, \tag{5.13}$$

where $\bar{\Lambda} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is defined as

$$\bar{\Lambda}(\bar{P}) = A'\bar{P}A - \bar{P} + (C' - A'\bar{P}\bar{C}')(\Lambda(0) - \bar{C}\bar{P}\bar{C}')^{-1}(C' - A'\bar{P}\bar{C}')'.$$

which is again the same function $\bar{\Lambda}$ introduced in the previous section.

The state $\hat{\bar{x}}(t)$ is the orthogonal projection onto $H_{[t,T]}$ (the minimum error-variance "filter" estimate) of the state $\bar{x}(t)$ of any minimal stationary backward realization of y in the uniform basis induced by $(A', \bar{C})$. As $T \to +\infty$ $\hat{\bar{x}}(t)$ converges to $\bar{x}_+(t)$ and the backward transient model tends to the steady state backward model,

$$\left\{ \begin{array}{rcl} \bar{x}_+(t-1) & = & A'\bar{x}_+(t) + \bar{B}_+\bar{w}_+(t-1) \\ y(t-1) & = & \bar{C}\bar{x}_+(t) + \bar{D}_+\bar{w}_+(t-1) \end{array} \right.$$

with state covariance $\bar{P}_+$.

In analogy to (5.7) the state $\hat{\bar{x}}(t)$ now spans the backward predictor space

$$\check{X}_{t+} := \text{E}^{H_{[t,T]}}H_t^-. \tag{5.14}$$

The backward form of the "forward Kalman Filter" and the "forward form" of the backward Kalman Filter. Exactly as it happens in the stationary setup, the joint processes $[\hat{x}(t)'\ y(t-1)']'$ and $[\hat{\bar{x}}(t)'\ y(t)]'$, being Markov processes, admit both a *causal* or *forward* and an *anticausal* or *backward* representation. Unfortunately the name of "backward Kalman Filter" given to the representation (5.11) refers to both circumstances that $\bar{x}(t)$ is a basis in the backward predictor space (i.e. is an estimate based on the future of y) and to the anticausal structure of the realization. These two facts are strictly speaking unrelated and, although it may look a bit unnatural, there is also a *causal* form of the backward Kalman Filter representation (5.11),

$$\begin{array}{rcl} \hat{\bar{x}}(t+1) & = & \bar{\Pi}(t+1)A\bar{\Pi}(t)^{-1}\hat{\bar{x}}(t) + \bar{B}_+(t)\epsilon_+(t) \\ y(t) & = & [\bar{C}\bar{\Pi}(t+1)A + \bar{D}(t)\bar{B}(t)']\bar{\Pi}(t)^{-1}\hat{\bar{x}}(t) + \bar{D}_+(t)\epsilon_+(t) \end{array} \tag{5.15}$$

where $\bar{B}_+(t), \bar{D}_+(t), \epsilon_+(t)$ have expressions which may be derived by the standard procedure explained e.g. in [48]. Of course the validity of the causal form above is subjected to the existence of the inverse $\bar{\Pi}(t)^{-1}$ which requires t "far enough" from the endpoint $t = T$. As shown in [48] and is visible in the last statement of Theorem 2.2, whenever we change causality structure it is convenient to go to the dual basis, namely define

$$\hat{x}_+(t) := \bar{\Pi}(t)^{-1}\hat{\bar{x}}(t) \tag{5.16}$$

and substitute into (5.15), which assume the simpler form

$$\begin{array}{rcl} \hat{x}_+(t+1) & = & A\hat{x}_+(t) + B_+(t)\epsilon_+(t) \\ y(t) & = & C\hat{x}_+(t) + D_+(t)\epsilon_+(t). \end{array} \tag{5.17}$$

The appearence of the C matrix in the output equation is due to the fact that

$$\begin{array}{rcl} [\bar{C}\bar{\Pi}(t+1)A + \bar{D}(t)\bar{B}(t)']\,\hat{x}_+(t) & = & \mathrm{E}[\,y(t)\,|\,\hat{x}_+(t)\,] = \\ \mathrm{E}(y(t)\hat{x}_+(t)')[\mathrm{E}\hat{x}_+(t)\hat{x}_+(t)']^{-1}\hat{x}_+(t) & = & \mathrm{E}(y(t)\hat{\bar{x}}(t)')\,\hat{x}_+(t), \end{array}$$

where, after inserting (5.10), the last member can be computed easily as

$$\begin{array}{rcl} \mathrm{E}y(t)(\mathrm{E}^{H_{[t,T]}}\bar{x}(t))' & = & \mathrm{E}y(t)\bar{x}(t)' = \mathrm{E}y(t)(A\bar{x}(t+1) + \bar{B}\bar{w}(t))' \\ & = & \bar{C}\bar{P}A' + \bar{D}\bar{B}' = C. \end{array}$$

Note the last identity is identical to (2.17).

5.2 Finite-interval realizations

The situation of interest is really when both the past and the future spaces are finitely generated. To discuss this context it is better to recall some general facts from the geometric theory.

Consider the class of *Markovian Splitting Subspaces* at time t, X_t, for the process y on the interval $[0,T]$. These subspaces make the joint finite past and future spaces conditionally orthogonal (i.e. uncorrelated) given X_t, i.e.,

$$H_{[0,t)} \vee X_{[0,t]} \perp H_{[t,T]} \vee X_{[t,T]} \big| X_t. \tag{5.18}$$

where the symbols have an obvious meaning.

It is standard [49, 50] to show that the forward and backward *finite-memory* predictor spaces,

$$\hat{X}_{t-} := \mathrm{E}^{H_{[0,t)}} H_{[t,T]} \quad \text{and} \quad \hat{X}_{t+} := \mathrm{E}^{H_{[t,T]}} H_{[0,t)},$$

are such Markovian splitting subspaces. In fact these predictor spaces are also *minimal splitting*, exactly as in the stationary theory. The following representation theorem (proven in [55]) relates the finite predictor spaces to Kalman filter realizations.

Theorem 5.1. *Let (2.1) be any minimal stationary realization of the process y and let X be the relative state space. Then*

$$X_t := U^t X, \quad 0 \le t \le T \tag{5.19}$$

is a Markovian splitting subspace for the process subordinated by y on the interval $[0,T]$ which is minimal if t is far enough from the endpoints of the interval. Here "far enough" means $t \ge \nu_c$ and $T - t \ge \nu_o$ where ν_c and ν_o are the reachability and observability indices of the pairs $(A, \bar{C}')$ and (A, C) respectively.

Under the same conditions on t and $T - t$, the finite memory predictor spaces $\hat{X}_{t-}$ and $\hat{X}_{t+}$ coincide with $\check{X}_{t-}$ and $\check{X}_{t+}$ and in fact we have the representation formulas

$$\hat{X}_{t-} = \mathrm{E}^{H_{[0,t)}} X_t \quad \textit{and} \quad \hat{X}_{t+} = \mathrm{E}^{H_{[t,T]}} X_t. \tag{5.20}$$

which hold for any stationary splitting subspace X.

One consequence of the representation formulas (5.19) is that for any choice of basis $x(0)$ in a minimal stationary state space X, the process

$$\hat{x}(t) := \mathrm{E}^{H_{[0,t)}} x(t), \quad t \in [0,T] \tag{5.21}$$

is the state process of a finite-interval realization of y which is minimal for all times t "far enough" from the endpoints of the interval. Of course the relative state equations, either in the forward or in the backward representation, are nothing else but the Kalman filter equations for $\{y(t)\}$, described previously. This in a sense is not big surprise.

What is more interesting for identification is the following converse statement. It says that every basis in a finite memory predictor space is the Kalman filter of a uniformly ordered family of minimal stationary realizations. Hence picking a basis in the finite memory predictor space is the same thing as picking a legitimate (A, C) pair of a minimal stationary realization of y.

Theorem 5.2. *Fix any minimal stationary splitting subspace X and assume that t is far enough from the endpoints of the interval $[0,T]$ in the sense explained before. Then, to a basis vector $\hat{x}(t)$ in $\hat{X}_{t-}$ there corresponds a* unique *basis $x(t)$ in X_t for which (5.21) holds. As X varies in the family of all minimal stationary splitting subspaces $\mathfrak{X}$ the bases $x(t)$ corresponding to a fixed $\hat{x}(t)$ describe a (forward) uniformly ordered family.*

Consequently a choice of basis in $\hat{X}_{t-}$ defines uniquely a matrix pair (A,C) of a minimal stationary realization of the process y.

A dual statement holds for the backward predictor space: any choice of basis $\hat{\bar{x}}(t)$ in $\hat{X}_{t+}$ defines uniquely a matrix pair $(\bar{A},\bar{C})$ corresponding to a uniform family of minimal backward stationary realization of y of which $\hat{\bar{x}}(t)$ is the Kalman Filter estimate given the finite future of y.

Proof. The statement follows from the representation (5.20) and the minimality of X_t as a splitting subspace for $H_{[0,t)}$ and $H_{[t,T]}$. Minimality in this sense in particular implies that the constructibility operator,

$$\mathcal{C}_t := \mathrm{E}^{H_{[0,t)}}_{|X_t} : X_t \to \hat{X}_{t-}$$

is *injective* [49]. In other words there are unique random variables $x_k(t) \in X_t$ which are projected onto the components $\hat{x}_k(t)$; $k = 1,\dots,n$.

The uniform order follows from the identity

$$\hat{x}(t) := \mathrm{E}^{H_{[0,t)}} x(t) = \mathrm{E}^{H_{[0,t)}} \mathrm{E}^{H_t^-} x(t).$$

It follows from this identity and from injectivity of $\mathcal{C}_t$ that all $x(t)$'s must have the same projection onto the infinite past $\mathrm{E}^{H_t^-} x(t)$. But this projection is the same as $\mathrm{E}^{X_t-} x(t)$ and so the bases $x(t)$ have the invariant projection property (4.11). The rest is immediate.

The following concept will play a role in identification.

Definition 5.1. *Two basis vectors $\hat{x}(t)$ in $\hat{X}_{t-}$ and $\hat{\bar{x}}(t)$ in $\hat{X}_{t+}$ are* coherent *if the corresponding (A,C) and $(\bar{A},\bar{C})$ pairs are such that $\bar{A} = A'$ and $(A,C,\bar{C},\Lambda(0))$ is a minimal realization of the spectral density matrix of y.*

Proposition 5.2. *Two basis vectors $\hat{x}(t)$ in $\hat{X}_{t-}$ and $\hat{\bar{x}}(t)$ in $\hat{X}_{t+}$ are coherent if and only if*

$$\hat{x}(t) := \mathrm{E}^{\hat{X}_{t-}} \hat{x}_+(t) \tag{5.22}$$

where $\hat{x}_+(t)$ is the dual basis of $\hat{\bar{x}}(t)$ defined in (5.16).

Proof. Consider the causal version, $\hat{x}_+(t)$, of the (unique) backward Kalman Filter corresponding to the uniform family of stationary backward models attached to the pair $(\bar{A},\bar{C})$ (or to $\hat{\bar{x}}(t)$). Clearly $\hat{\bar{x}}(t)$ is coherent with $\hat{x}(t)$ if and only if this causal nonstationary model has a time-invariant (A,C) pair coincident with the (A,C) pair of $\hat{x}(t)$. But then by the uniqueness theorem 5.2 $\hat{x}_+(t)$ must coincide with $\hat{x}(t)$.

Remark. As we shall see later, the main idea of "subspace methods" identification is to recapture the *stationary* A, C and $\bar{A}, \bar{C}$ parameters of the process from the dynamic equations satisfied by the bases $\hat{x}(t)$ and $\hat{\bar{x}}(t)$ chosen in the finite-memory predictor spaces. As we have seen before these equations can be written as the Kalman Filter realizations (5.5, 5.6) or 5.11) where A, C and $\bar{A}, \bar{C}$ appear explicitly. However it should be stressed that the stationary parameters A, C and $\bar{A}, \bar{C}$ appear in the Kalman-Filter equations exactly because the bases $\hat{x}(t)$ and $\hat{\bar{x}}(t)$ have been obtained by projection of some stationary state $x(t)$ of the process. In identification, where we are actually attempting to recover the stationary dynamics of y, we do not have a stationary state-space model for y at our disposal. Instead we can pick the bases in the predictor spaces $\hat{X}_{t-}, \hat{X}_{(t+1)-}, \ldots$ at different time instants and compute the difference equations relating the bases at different time instants.

Of course one could in principle pick a basis arbitrarily at different time instants but this would naturally yield *time-varying* A and C parameters in the state equations. In fact, picking bases arbitrarily at each instant t yields time-varying A and C matrices which are not even similar to the stationary parameters we are looking for. So the question arises of choosing bases $\hat{x}(t)$ and $\hat{\bar{x}}(t)$ at different instants $t \in [0, T]$ in such a way that their time evolution is of the Kalman Filter type encountered so far, in particular described by difference equations with *constant* matrices A and C. In this case, by Theorem 5.2, A and C will be equal to the corresponding parameters of a stationary model (in fact, of the whole uniformly ordered family of stationary models corresponding to the basis, see Theorem 5.2).

Now the Kalman Filter equations describe the propagation in time of the projection $\hat{x}(t)$ onto the past $H_{[0,t)}$, of a stationarily time-shifted state variable $x(t) = U^t x(0)$. Recall that the maps,

$$\mathcal{C}_t := \mathrm{E}^{H_{[0,t)}}_{|X_t} : X_t \to \hat{X}_{t-}, \quad \mathcal{O}_t := \mathrm{E}^{H_{[t,T]}}_{|X_t} : X_t \to \hat{X}_{t+}$$

are the *constructibility and observability operators* of X_t [49, 50] and that minimality of X_t implies that $\mathcal{C}_t, \mathcal{O}_t$ are invertible in their respective codomains (we have used this argument already in the proof of Theorem 5.2 above). We may define then a forward and backward *conditional shift operator* $\hat{U}(t)$ and $\hat{\bar{U}}(t)$, by setting

$$\hat{U}(t) := \mathcal{C}_{t+1} U \mathcal{C}_t^{-1} : \hat{X}_{t-} \to \hat{X}_{(t+1)-} \tag{5.23}$$

and

$$\hat{\bar{U}}(t) := \mathcal{O}_{t-1} U^{-1} \mathcal{O}_t^{-1} : \hat{X}_{t+} \to \hat{X}_{(t-1)+} \tag{5.24}$$

so that

$$\begin{aligned} \hat{x}(t+1) = \mathrm{E}^{H_{[0,t+1)}} x(t+1) &= \mathrm{E}^{H_{[0,t+1)}} U\, x(t) = \hat{U}(t)\hat{x}(t) \\ \hat{\bar{x}}(t-1) = \mathrm{E}^{H_{[t-1,T]}} \bar{x}(t-1) &= \mathrm{E}^{H_{[t-1,T]}} U^{-1}\, \bar{x}(t) = \hat{\bar{U}}(t)\hat{\bar{x}}(t). \end{aligned}$$

It is not difficult to show (but we shall not do it here) that the definition of $\hat{U}$ and $\hat{\bar{U}}$ is independent of the minimal stationary splitting subspace X entering in the equations (5.23, 5.24). This specific form of propagation in time, say for the basis $\hat{x}(t)$, is equivalent to the relative state equations being of the Kalman filter type, in particular involving A and C matrices which stay constant in time.

Proposition 5.3. *Let $\hat{x}(t)$ be a basis in $\hat{X}_{t-}$ and $\hat{z}(t+1)$ be an arbitrary basis in $\hat{X}_{(t+1)-}$. Then $\hat{z}(t+1)$ is the conditional shift of $\hat{x}(t)$, i.e. $\hat{z}(t+1) = \hat{x}(t+1) = \hat{U}(t)\hat{x}(t)$ if and only if*

$$\mathrm{E}[\hat{z}(t+1)|\,\hat{x}(t)\,] = A\hat{x}(t), \quad \mathrm{E}y(t)\hat{z}(t+1)' = \bar{C}$$

where A is the state-transition matrix of the uniform choice of bases determined by $\hat{x}(t)$ and $\bar{C}$ is the corresponding backward state-output matrix.

Dually, let $\hat{\bar{x}}(t)$ be a basis in $\hat{X}_{t+}$ and $\hat{\bar{z}}(t-1)$ be an arbitrary basis in $\hat{X}_{(t-1)+}$. Then $\hat{\bar{z}}(t-1)$ is the conditional shift of $\hat{\bar{x}}(t)$, i.e. $\hat{\bar{z}}(t-1) = \hat{\bar{x}}(t-1) = \hat{\bar{U}}(t)\hat{\bar{x}}(t)$ if and only if

$$\mathrm{E}[\hat{\bar{z}}(t-1)|\,\hat{\bar{x}}(t)\,] = A'\hat{\bar{x}}(t), \quad \mathrm{E}y(t)\hat{\bar{z}}(t-1)' = C$$

where A' is the state-transition matrix of the (backward) uniform choice of bases determined by $\hat{\bar{x}}(t)$ and C is the corresponding forward state-output matrix.

Proof. (if). By the Markovian splitting property every family of bases $\{\hat{z}(t); t \in [0,T]\}$ in the predictor spaces $\{\hat{X}_{t-}; t \in [0,T]\}$ forms a Markov process. It is also easy to see that the past space $\mathrm{span}\{\hat{z}(s); s < t\}$ coincides with $H_{[0,t)}$ for all $t \geq 0$. Hence

$$\mathrm{E}[\hat{z}(t+1)|\,H_{[0,t)}\,] = \mathrm{E}[\hat{z}(t+1)|\,\hat{x}(t)\,] = A\hat{x}(t)$$

which is by assumption equal to

$$\mathrm{E}[\hat{x}(t+1)|\,\hat{x}(t)\,] = \mathrm{E}[\hat{x}(t+1)|\,H_{[0,t)}\,].$$

Therefore $\hat{z}(t+1)$ and $\hat{x}(t+1)$ have the same orthogonal projection onto $H_{[0,t)}$ i.e.

$$\mathrm{E}^{H_{[0,t)}}(\hat{z}(t+1) - \hat{x}(t+1)) = 0$$

so that, letting $Y_t^- := [y(t)'y(t-1)'\dots y(0)']'$ and taking into account also the second condition in the statement of the proposition involving $\bar{C} = \mathrm{E}y(t)\hat{x}(t+1)'$ (this identity follows since $\bar{C}\hat{\bar{x}}_-(t+1) = \mathrm{E}[y(t)\,|\hat{\bar{x}}_-(t+1)\,]$ where $\hat{\bar{x}}_-(t+1)$ is the dual basis of $\hat{x}(t+1)$), we have

$$\mathrm{E}Y_t^-\hat{z}(t+1)' = \mathrm{E}Y_t^-\hat{x}(t+1)' = \begin{bmatrix} \bar{C} \\ \bar{C}A' \\ \vdots \\ \bar{C}(A')^t \end{bmatrix}.$$

We denote the constructibilty matrix on the right by $\bar{\Omega}$. Now clearly $\hat{z}(t+1) = M\hat{x}(t+1)$ for some nonsingular matrix M and by substituting in the first member of the equality yields $\bar{\Omega}M' = \bar{\Omega}$ which is equivalent to $M = I$ since (for t large enough) $\bar{\Omega}$ has independent columns. So $\hat{z}(t+1)$ coincides with the conditionally shifted basis $\hat{x}(t+1)$.

(only if). Pick any $x(t)$ in the stationary uniform family of bases corresponding to $\hat{x}(t)$ ($x(t) = \mathcal{C}_t^{-1}\hat{x}(t)$, see Theorem 5.2). Then

$$\hat{x}(t+1) = \hat{U}(t)\hat{x}(t) = \mathrm{E}[x(t+1)|\, H_{[0,t+1)}\,]$$

is precisely the Kalman Filter estimate of the stationary state process x at time $t+1$. So the formulas follow from the derivations of the Kalman Filter equations at the beginning of the section.

The proof of the second half of the proposition involves completely similar arguments and is skipped.

6. Estimation, partial realization and balancing

We shall now concentrate on the statistical problem of describing an observed m-dimensional time series

$$\{y_0, y_1, y_2, \dots\, y_T\}, \tag{6.1}$$

by a finite-dimensional state-space model of the type (2.1) studied in the previous sections.

To put the methods discussed in this paper into perspective we should say that there are also different choices of the model class which are widely used in identification. One may choose,

1. A parametric class of spectral density functions; say all the rational spectra $\Phi(z)$ of fixed McMillan degree n.
2. A parametric class of (rational) minimal *shaping filter* representations, in other words models consisting of a pair: minimal spectral factor W, plus input white noise w. Expressing W as a polynomial matrix fraction,

   $$W(z) = A(z^{-1})^{-1}B(z^{-1})$$

   gives the model the familiar form of a linear difference equation

   $$y(t) + \sum_{k=1}^{\nu} A_k y(t-k) = \sum_{k=0}^{\nu} B_k w(t-k) \tag{6.2}$$

   i.e. an "ARMA" model, parametrized by the coefficients $\{A_k, B_k\}$ of the matrix polynomials ($A(z^{-1})$, $B(z^{-1})$). As we have seen at the end of section 4, for square W's the input noise is uniquely determined by the output signal y.

3. Minimal state-space realizations of the type (2.1). These objects are the most "structured" kind of representation of the signal and can be reduced to the previous kind of models by eliminating the auxiliary variables (x and w). They will be our primary object of interest.

For each model class there is a problem of *unique parametrization*, i.e. of making the correspondence: parameter $\rightarrow$ model, generically bijective. The solution of this problem via the theory of canonical forms constitutes an important chapter of identification theory which has attracted much interest in the early seventies but is now a bit obsolete since *balanced canonical forms* [58], [59], which will be introduced later, are a much simpler and robust alternative.

Moreover, while a spectral density is a unique (wide-sense) probabilistic description of a signal, a family of different minimal spectral factors or state-space models (neglecting the indeterminacy inherent in the choice of basis) give rise to the same spectrum. For this reason when the model classes (2) and (3) are used it is necessary to specify a *representative* factor or minimal realization to get a 1:1 correspondence with the spectrum. Normally one chooses to describe a spectrum by its (unique) minimum phase spectral factor or *forward innovation models* i.e. or the corresponding causal "steady state Kalman Filter" realization. These models are 1:1 with the spectrum if we disregard the intrinsic indeterminacy in the input white noise (which is only defined modulo constant real orthogonal transformations) and the arbitrariness in the choice of basis in the relative state space X_-.

The model classes described above are wide-sense. In case the signal y is believed to be *Gaussian* they can equivalently be interpreted as defining the spectrum or the covariance function of a family of Gaussian probability laws for the underlying stochastic process. These probabiliy laws are uniquely determined by a corresponding model and are then also parametrized by the parameters $\{A, C, \bar{C}, \Lambda(0)\}$, $\{A_k, B_k\}$ and (A, B, C, D) respectively.

We shall consider two conceptually different approaches that are used to fit models to the data,

– The "direct" approach, based on the principle of minimizing a suitable function which measures the distance between the data [13] and the probability law induced by the model class. Well-known and widely accepted examples of distance functions are the *likelihood function* of the data according to the particular model, or the average squared *prediction-error* of the observed data corresponding to a particular choice of a model in the model class. Minimization of these criteria can (except in trivial cases) only

[13] This terminology is a bit misleading. In reality one minimizes a suitable "finite sample" approximation of a distance function between the *true law* of the data and the law induced by the model class. An example of distance function between probability measures which can be used to this purposse is the Kullback-Leibler distance.

be done numerically and hence the direct methods lead to iterative optimization algorithms in the space of the parameters, say the space of minimal (A, B, C, D) matrix quadruples, which parametrize the chosen model class.

– A two steps procedure which in principle can be described as identification of a rational model for the spectrum (or covariance) of the observed data, followed by stochastic realization. Here the first step is estimation of the parameters $(A, C, \bar{C})$ of a minimal realization of the spectral density matrix of the process. From the spectral density matrix a state-space model (typically the forward innovation model) is then computed by solving the Linear Matrix Inequality, or the Riccati equation as seen in section 4.
The difference with the first approach is that the estimation of $(A, C, \bar{C})$ is *not* done by optimizing a likelihood or other distance functions but simply by *matching second order moments*. In other words, let

$$\{\Lambda_0, \Lambda_1, \ldots, \Lambda_\nu\} \tag{6.3}$$

be a finite set of sample $m \times m$ covariance matrices estimated in some (as yet unspecified) way from the m-dimensional sequence of observations (6.1). The problem is of finding a minimal value of n and a minimal[14] triplet of matrices $(A, C, \bar{C})$, of dimensions $n \times n$, $m \times n$ and $m \times n$ respectively, such that

$$CA^{k-1}\bar{C}' = \Lambda_k \quad k = 1, 2, \ldots, \nu \tag{6.4}$$

This is an instance of *estimation by the method of moments* described in the statistical textbooks e.g. [13, p. 497], which is a very old idea used extensively by K. Pearson in the beginning of the century. The underlying priciple is close in spirit to the wide-sense setting that we are working in. It does not necessarily guarantee minimal distance between the "true" and the model distributions but rather imposes that the parameters to be estimated match exactly the sample second order moments. These can easily be chosen at least "consistent" (i.e. tending to the true second order moments as the sample size goes to infinity) so the method gives consistent estimates in the sense that ν true moments $\Lambda_0(\tau) \quad \tau = 1, 2, \ldots, \nu$ will be described exactly as $T \to \infty$. In other words the first ν lag values of the true covariance function will be matched exactly.

Some may argue that estimation by the method of moments is in general "non-efficient" and it is generally claimed in the literature that one should expect better results (in the sense of smaller asymptotic variance of the estimates) by direct methods. In practice this is true only to a point since the likelihood function or the average prediction error are computable only if we assume Gaussian models (or linear predictors which amounts to the same)

[14] Recall that $(A, C, \bar{C})$ is minimal if (A, C) is completely observable and $(A, \bar{C}')$ is completely reachable.

and this in the long run is equivalent to matching covariances anyway. In addition there is the structural handicap of iterative optimization methods which may get stuck in local minima and hence provide sub-optimal parameter estimates, a rather hard phenomenon to detect. The two-steps approach offers in this respect the major advantage of converting the nonlinear parameter estimation phase which is necessary in maximum-likelihood or prediction-error model identification into a partial realization problem, involving essentially the factorization of a Hankel matrix of estimated covariances, and the solution of a Riccati equation, both much better understood problems for which efficient numerical solution techniques are available.

6.1 Positivity

A warning is in order concerning the implementation of the method of moments described above in that it introduces some nontrivial mathematical questions related to positivity of the estimated spectrum.

In determining a minimal triplet $(A, C, \bar{C})$ interpolating the partial sequence (6.3) so that $CA^{k-1}\bar{C}' = \Lambda_k \quad k = 1, 2, \dots, \nu$, we also completely determine the infinite sequence

$$\{\Lambda_0, \Lambda_1, \Lambda_2, \Lambda_3, \dots\} \tag{6.5}$$

by setting $\Lambda_k = CA^{k-1}\bar{C}'$ for $k = \nu + 1, \nu + 2, \dots$. This sequence is called a *minimal rational extension* of the finite sequence (6.3). The attribute "rational" is due to the fact that

$$Z(z) := \frac{1}{2}\Lambda_0 + \Lambda_1 z^{-1} + \Lambda_2 z^{-2} + \dots = \frac{1}{2}\Lambda_0 + C(zI - A)^{-1}\bar{C}' \tag{6.6}$$

is a rational function. In order for (6.5) to be a bona fide covariance sequence, however, it is necessary, but *not* sufficient, that the Toeplitz matrix

$$T = \begin{bmatrix} \Lambda_0 & \Lambda_1 & \Lambda_2 & \cdots & \Lambda_\nu \\ \Lambda_1' & \Lambda_0 & \Lambda_1 & \cdots & \Lambda_{\nu_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_\nu' & \Lambda_{\nu-1}' & \Lambda_{\nu-2} & \dots & \Lambda_0 \end{bmatrix} \tag{6.7}$$

be nonnegative definite. In fact, it is required that the function (the spectral density corresponding to (6.5))

$$\Phi(z) = \Lambda_0 + \sum_{k=1}^{\infty} \Lambda_k (z^k + z^{-k}) = Z(z) + Z(z^{-1})' \tag{6.8}$$

be nonnegative on the unit circle. This is equivalent to the function $Z(z)$ being *positive real*. Consequently, the interpolation needs to be done subject to the extra constraint of positivity.

The constraint of positivity is a rather tricky one and in all identification methods which are directly or indirectly, as the subspace methods described below, based on the interpolation condition (6.4) it is normally disregarded. For this reason these methods may fail to provide a positive extension and hence may lead to data $(A, C, \bar{C})$ for which there are no solutions of the LMI and hence to totally inconsistent results.

It is important to appreciate the fact that the problem of positivity of the extension has little to do with the "noise" or "sample variability" superimposed to the covariance data and is present equally well for (finite) data extracted from a true rational covariance sequence. For there is no guarantee that, even in this idealized situation, the order of a minimal rational extension 6.5 of the first ν covariance matrices of the sequaence, would be sufficiently high to equal the order of the infinite sequence and hence to generate a positive extension. A minimal partial realization may well fail to be positive because its order is too low to guarantee positivity.

Neglecting the positivity constraint amounts to tacitly assuming that

Assumption 6.1. The covariance data (6.3) can be generated exactly by some (unknown) stochastic system whose dimension is equal to the rank of the block Hankel matrix

$$H_\mu = \begin{bmatrix} \Lambda_1 & \Lambda_2 & \Lambda_3 & \cdots & \Lambda_\mu \\ \Lambda_2 & \Lambda_3 & \Lambda_4 & \cdots & \Lambda_{\mu+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_\mu & \Lambda_{\mu+1} & \Lambda_{\mu+2} & \cdots & \Lambda_{2\mu-1} \end{bmatrix}, \tag{6.9}$$

where $\mu = [\frac{\nu}{2}]$.

This assumption is not "generically satisfied" and it can be shown that there are relatively "large" sets of data (6.3) for which it does not hold. It is not even enough to assume that the data is generated from a "true" finite-dimensional stochastic system: the rank condition is also necessary. Otherwise, for a minimal triplet $(A, C, \bar{C})$ which satisfies the interpolation condition (6.4), the positivity condition will not be automatically fulfilled, and the matrix A may even fail to be stable [10].

Following [55], we define the *algebraic degree* of the sequence (6.3) to be the minimal degree of any realization (**??**) satisfying (6.4) and the *positive degree* to be the minimal degree of a rational extension (**??**) for which, in addition, $Z(z)$ is positive real. Then Assumption 6.1 can also be written in the following equivalent form.

Assumption 6.1′. The positive degree of (6.3) is equal to the algebraic degree.

The fact that this equality cannot a priori be assumed to hold for generic covariance data can now be illustrated by the following fact.

Theorem 6.1. *[11] In the case $m = 1$, the generic value of the algebraic degree of (6.3) is $[\frac{\nu+1}{2}]$, whereas there is no generic value for the positive degree. In fact, for each $p = [\frac{\nu+1}{2}], [\frac{\nu+1}{2}] + 1, \dots, \nu$, there is a nonempty open set of covariance data for which the positive degree is precisely p.*

The correct approach would in principle require to compute a rational *positive extension* of the finite covariance sequence (6.3), of minimal McMillan degree. Although there are methods to compute positive extensions, the most famous of which is the so-called "maximum-entropy" extension, based on the Levinson algorthm, these methods produce functions of very high complexity, in fact generically of the highest possible degree (ν in the case $m = 1$). Unfortunately there are no algorithms so far which compute positive extensions of minimal degree. A stochastic model reduction step would then be necessary but this is again, a rather underdeveloped area of system theory. For a discussion of these matters see [55].

In these notes we shall be content with discussing the deterministic partial realization aspect of the method, therefore tacitly assuming that the conditions described in Assumption 6.1 hold. There is standard software available for checking positivity (i.e. solvability of the LMI) of the partial realization. Whenever positivity fails one may try to add more covariance data so as to allow for an increase of the order (algebraic degree) of the partial realization. Once a positive triple $(A, C, \bar{C})$ is estimated, the computation of a state-space model is in principle just a matter of solving the LMI or the appropriate Riccati equation, as seen in section 4..

Historical remarks. The two-steps procedure was apparently first advocated in a systematic way by Faurre [21]; see also [22, 23]. More recent work is based on Singular Value Decomposition and canonical correlation analysis [2] and is due to Aoki [9], and van Overschee and De Moor [60]. There are versions of the algorithms based on canonical correlation analysis which apply directly to the observed data without even computing the covariance estimates [60].

The work of van Overschee and De Moor introduces an interesting "geometric" approach based on state-space construction and on the choice of particular bases in the state space. The system matrices are computed after the choice of basis by formulas analog to (2.12). This procedure on one hand makes very close contact with the geometric state-space construction ideas discussed in sections5.2 and 2. On the other hand it seems completely unrelated to the partial realization and covariance extension approach mentioned above.

In the rest of this paper we shall study the geometric "Subspace-methods" approach of [60] and show that it is very much related to the basic partial realization plus stochastic realization idea. In fact we shall show that the two approaches are equivalent and lead to exactly the same formulas.

6.2 The Hilbert Space of a Stationary signal

In section 2. we have described an abstract model-building procedure based on geometric operations on certain subspaces of random variables constructed from linear statistics of the present and past histories of a stochastic process y. In practice instead one has just a collection of observed data,

$$\{\, y_0, y_1, \ldots, y_t, \ldots, y_T \,\} \tag{6.10}$$

with $y_t \in \mathbb{R}^m$, measured during an experiment. We shall assume that the sample size T is very large and that the data have been preprocessed so as to be compatible with the basic assumption of (wide-sense) stationarity and zero mean of the previous sections. This in particular means that we can pick N large enough so that the time averages

$$\frac{1}{N+1} \sum_{t=t_0}^{N+t_0} y_{t+\tau} y_t' \qquad \tau \geq 0 \tag{6.11}$$

are practically independent of the initial time t_0 and arbitrarily close to a *bona-fide* stationary covariance matrix sequence (the "true" covariance of the signal).

Under these assumptions on the data, the stochastic state-space theory of the sections 2.-5. can be translated into an isomorphic geometrical setup based on linear operations on the observed time series and can then applied to the problem of state-space modeling of the data.

In this section we shall briefly review the basic ideas behind this correspondence. For clarity of exposition we shall initially assume that $T = \infty$ and that the data collection has started in the infinitely remote past (so that the time series is actually doubly infinite).

For each $t \in \mathbb{Z}$ define the $m \times \infty$ matrices

$$\mathbf{y}(t) := [y_t, y_{t+1}, y_{t+2}, \ldots] \tag{6.12}$$

and consider the sequences $\mathbf{y} := \{\mathbf{y}(t) \,|\, t \in \mathbb{Z}\}$. This sequence will play a very similar role to the stationary processes y of the previous sections.

Define the vector space $\mathcal{Y}$ of all finite linear combinations

$$\mathcal{Y} := \{\sum a_k' \mathbf{y}(t_k) \qquad a_k \in \mathbb{R}^m,\ t_k \in \mathbb{Z}\} \tag{6.13}$$

Note that the vector space $\mathcal{Y}$ is just the row spaces of the family of semi-infinite matrices (6.12) or, equivalently the rowspace of the infinite Hankel matrix

$$Y_\infty := \begin{bmatrix} \vdots \\ \mathbf{y}(t) \\ \mathbf{y}(t+1) \\ \mathbf{y}(t+2) \\ \vdots \end{bmatrix}$$

This vector space of scalar semi-infinite sequences (rows) can be equipped with an inner product, which is first defined on the generators by the bilinear form

$$\langle a'\mathbf{y}(k), b'\mathbf{y}(j)\rangle := \lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} a'y_{t+k}y'_{t+j}b = a'\Lambda_0(k-j)b, \qquad (6.14)$$

and then extended by linearity to all finte linear combinations of elements of $\mathcal{Y}$. This inner product is nondegenerate if the Toeplitz matrix T_k, constructed with the true covariances $\{\Lambda_0(0), \Lambda_0(1), \dots, \Lambda_0(k)\}$, is a positive definite symmetric matrix for all k [55]. Note also that the limit does not change if in the limits of the sum (6.14) $t = 0$ is replaced by an arbitrary initial instant t_0, so that

$$\langle a'\mathbf{y}(k), b'\mathbf{y}(j)\rangle = \langle a'\mathbf{y}(t_0+k), b'\mathbf{y}(t_0+j)\rangle$$

for all t_0 (wide-sense stationarity). We also define a *shift operator* $\mathbf{U}$ on the family of semi-infinite matrices (6.12), by setting

$$\mathbf{U}a'\mathbf{y}(t) := a'\mathbf{y}(t+1) \quad t \in \mathbb{Z}, \quad a \in \mathbb{R}^m,$$

defining a linear map which is isometric with respect to the inner product (6.14) and extendable by linearity to all of $\mathcal{Y}$.

By closing the vector space $\mathcal{Y}$ with respect to convergence in the norm induced by the inner product (6.14), we obtain a Hilbert space [15] $\bar{\mathcal{Y}} = \text{closure}\{\mathcal{Y}\}$ to which the shift operator $\mathbf{U}$ is extended by continuity as a unitary operator.

As explained in more detail in [55], this Hilbert space framework is isometrically isomorphic to the abstract "stochastic" geometric setup used in the previous sections. Now as stated formally in Proposition1.1 and in the subsequent generalization, we can formally think of the observed (infinitely long) time series as a regular sample path of a wide-sense stationary stochastic process $\mathbf{y}$, having covariance matrix equal to the true covariance function $\Lambda_0(.)$, equal to the limit of the sum (6.11) as $N \to \infty$. Then, at least as far as first and second order moments are concerned, the sequence of "tails" $\mathbf{y}$ defined in (6.12) behaves exactly like the abstract stochastic counterpart y. In particular all second order moments of the random process can equivalently be calculated in terms of the tail sequence $\mathbf{y}$ provided we substitute expectations with ergodic limits of the type (6.14). Since we only worry about second order properties in this paper, we may even formally *identify* the tail sequence $\mathbf{y}$ of (6.12) with the underlying stochastic process y. This requires just thinking of "random variables" as being semi-infinte strings of numbers and the expectation of products $\mathrm{E}\{\xi\eta\}$ as being the inner product of the

[15] Note that the symbol $\mathcal{Y}$ denotes a real inner-product space which need not be closed with respect to the inner product structure defined by (6.14). Since we will not have much use for the completed space in the following, we shall not introduce special symbols for it.

corresponding rows ξ and η. For reasons of uniformity of notation the inner product 6.14 will then be denoted

$$\langle \xi, \eta \rangle = \mathrm{E}\{\xi\eta\}, \tag{6.15}$$

Here as usual we allow $\mathrm{E}\{\cdot\}$ to operate on matrices, taking inner products row by row.

Hence all defintions and results in the geometric theory of stochastic realization can be carried over to the present framework. The orthogonal projection of ξ onto a subspace $\mathcal{H}$ of the space $\mathcal{Y}$ will still be denoted $\mathrm{E}[\,\xi\,|\mathcal{H}\,]$. Whenever $\mathcal{H}$ is given as the rowspace of some matrix of generators H, we shall write $\mathrm{E}[\,\xi\,|\,H\,]$ to denote the projection expressed (perhaps nonuniquely) in terms of the generators. It is clear that for finitely generated subspaces we have the representation formula

$$\mathrm{E}[\,\xi\,|\,H\,] = \mathrm{E}(\xi H')[\mathrm{E}(HH')]^\sharp H \tag{6.16}$$

and in case of linearly independent rows we can substitute the pseudoinverse $\sharp$ with a true inverse.

A (stationary) stochastic realization of $\mathbf{y}$ is a representation of the type

$$\left\{ \begin{array}{lcl} \mathbf{x}(t+1) & = & A\mathbf{x}(t) + B\mathbf{w}(t) \\ \mathbf{y}(t) & = & C\mathbf{x}(t) + D\mathbf{w}(t) \end{array} \right. \tag{6.17}$$

where $\{\mathbf{w}(t)\}$ is p-dimensional normalized white noise , i.e. $\mathrm{E}\{\mathbf{w}(t)\mathbf{w}(s)'\} = I\delta_{ts}$ $\mathrm{E}\{\mathbf{w}(t)\} = 0$, etc.

Remark 6.1. It should be kept in mind that the various linear operations in (6.17) hold in the sense of the metric of the space $\bar{\mathcal{Y}}$ defined above and are to be understood as "asymptotic equalities" between *sequences*. In particular, nothing can be said about the particular sample values, say y_t, x_t, w_t taken on by the time series involved in the model at a specific instant of time. This is similar to the interpretation that is given to the model (2.1) in case of *bona fide* stochastic processes, where the linear model can be expected to hold for each particular sample value only with probability one.

6.3 Identification based on Finite Data

For data of finite lenght T the inner product (6.15) must be approximated by a finite sum

$$E\{\xi\eta\} \cong \frac{1}{T+1} \sum_{t=0}^{T} \xi_t \eta_t \tag{6.18}$$

which makes the "expectation" operator E essentially the same thing as ordinary Euclidean inner product in $\mathbb{R}^T$.

Assume $N \leq T$ is large enough for the time average in the ergodic limit (6.11) to be sufficiently close to the true covariance and for all subscripts below to make sense. Fix a "present" time $t = k$ and define the two mk-dimensional "random vectors" (i.e. block Hankel matrices of dimension $mk \times (N+1)$) formed by stacking the output data as

$$\mathbf{Y}_k^- = \begin{bmatrix} \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \\ \mathbf{y}(k-1) \end{bmatrix} = \begin{bmatrix} y_0 & y_1 & \cdots & y_N \\ y_1 & y_2 & \cdots & y_{N+1} \\ \vdots & \vdots & & \vdots \\ y_{k-1} & y_k & \cdots & y_{k+N-1} \end{bmatrix} \tag{6.19}$$

$$\mathbf{Y}_k^+ = \begin{bmatrix} \mathbf{y}(k) \\ \mathbf{y}(k+1) \\ \vdots \\ \mathbf{y}(2k-1) \end{bmatrix} = \begin{bmatrix} y_k & y_{k+1} & \cdots & y_{k+N} \\ y_{k+1} & y_{k+2} & \cdots & y_{k+N+1} \\ \vdots & \vdots & & \vdots \\ y_{2k-1} & y_{2k} & \cdots & y_{2k+N-1} \end{bmatrix} \tag{6.20}$$

The relative rowspaces $\mathbf{Y}_k^-$, $\mathbf{Y}_k^+$ generated by the rows of the $m \times (N+1)$ matrices $\mathbf{y}(t)$ for $0 \leq t < k$, and $k \leq t < 2k$ respectively, are the "past" and "future" spaces of the data at time k. Since the tail matrix sequences we can form with the observed signal are necessarily finite, these vector spaces can describe in reality only *finite* past and future histories of the signal $\mathbf{y}$ at time k. For simplicity of notations we use symbols that are not informative of this fact[16].

For later use let us define also the "augmented" future at time k (a $m(k+1) \times (N+1)$ block Hankel matrix)

$$\mathbf{Y}_{[k,2k]}^+ := \begin{bmatrix} \mathbf{Y}_k^+ \\ \mathbf{y}(2k) \end{bmatrix},$$

the relative rowspace will be denoted $\mathbf{Y}_{[k,2k]}$.

6.4 The partial realization problem

In order to avoid trivial difficulties having to do with the fact that the rank of a finite Hankel matrix with too few rows or columns need not be equal to the algebraic degree of a finite sequence (6.3), we shall assume that the index k is chosen far enough from the endpoints $k = 0$ or $k = \nu$. There is in fact no loss of generality in assuming that we have $\nu = 2k+1$ sample covariance estimates,

$$\{\Lambda_0, \Lambda_1, \ldots, \Lambda_{2k}\}. \tag{6.21}$$

[16] More accurate notations would be,

$$\mathbf{Y}_k^- := \mathbf{Y}_{[0,k)} \quad \mathbf{Y}_k^+ := \mathbf{Y}_{[k,2k)}$$

and that the present time k has been chosen to be the "middle point" of the lag sequence of the covariance estimates (6.21).

A block Hankel matrix of stationary covariances can always be given the meaning of cross covariance matrix of the finite future and past of the underlying signal at time $t = k$,

$$\begin{aligned} H_k &:= \begin{bmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_k \\ \Lambda_2 & \Lambda_3 & \cdots & \Lambda_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_k & \Lambda_{k+1} & \cdots & \Lambda_{2k-1} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} \mathbf{y}(k) \\ \mathbf{y}(k+1) \\ \vdots \\ \mathbf{y}(2k-1) \end{bmatrix} \begin{bmatrix} \mathbf{y}(k-1) \\ \mathbf{y}(k-2) \\ \vdots \\ \mathbf{y}(0) \end{bmatrix}' \\ &= \mathrm{E}\mathbf{Y}_k^+(\bar{\mathbf{Y}}_k^-)' \end{aligned} \tag{6.22}$$

where $\bar{\mathbf{Y}}_k^-$ is the "time reversal" of the vector $\mathbf{Y}_k^-$. The subscript k is attached to denote the "present" time. Similarly using the available covariance data we can form

$$\begin{aligned} H_{k+1} &:= \begin{bmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_k & \Lambda_{k+1} \\ \Lambda_2 & \Lambda_3 & \cdots & \Lambda_{k+1} & \Lambda_{k+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_k & \Lambda_{k+1} & \cdots & \Lambda_{2k-1} & \Lambda_{2k} \end{bmatrix} = \begin{bmatrix} \Lambda_1 & & \\ \Lambda_2 & & \\ \vdots & & \sigma H_k \\ \Lambda_k & & \end{bmatrix} \\ &= \mathrm{E}\{U\mathbf{Y}_k^+(\bar{\mathbf{Y}}_{k+1}^-)'\}, \end{aligned} \tag{6.23}$$

and

$$\begin{aligned} \bar{H}_{k+1} &:= \begin{bmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_k \\ \Lambda_2 & \Lambda_3 & \cdots & \Lambda_{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_k & \Lambda_{k+1} & \cdots & \Lambda_{2k-1} \\ \Lambda_{k+1} & \Lambda_{k+2} & \cdots & \Lambda_{2k} \end{bmatrix} = \begin{bmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_k \\ & \sigma H_k & & \end{bmatrix} \\ &= \mathrm{E}\{\mathbf{Y}_{[k,2k]}^+(\bar{\mathbf{Y}}_k^-)'\} \end{aligned} \tag{6.24}$$

where σH_k is the *shifted Hankel matrix*, of the same dimension as H_k but with all entries shifted by one time unit i.e. with Λ_{i+1} replacing Λ_i everywhere.

We quote the following uniqueness result of partial realizations from [69].

Lemma 6.1. *The sequence (6.21) has a unique rational extension of minimal degree if and only if*

$$\mathrm{rank}H_k = \mathrm{rank}H_{k+1} = \mathrm{rank}\bar{H}_{k+1} := n \tag{6.25}$$

Uniqueness is understood in the sense that if $(A_1, C_1, \bar{C}_1)$ and $(A_2, C_2, \bar{C}_2)$ both define minimal rational extensions of (6.21), then there is a nonsingular $n \times n$ matrix T such that

$$A_2 = T^{-1}A_1T, \quad C_2 = C_1T, \quad \bar{C}_2' = T^{-1}\bar{C}_1'. \tag{6.26}$$

Computing a minimal partial realization can be done essentially via a rank factorization of the Hankel matrix H_k. The prototype algorithm, called the *Ho-Kalman* algorithm is reviwed below.

The Ho-Kalman Algorithm. Start by a *rank factorization* of H_k,

$$H_k = \Omega_k \bar{\Omega}_k' \tag{6.27}$$

where both factors Ω_k, $\bar{\Omega}_k$ have n linearly independent columns. Since by (6.25) column $-$ spanH_k = column $-$ spanH_{k+1} and, dually, row $-$ spanH_k = row $-$ span$\bar{H}_{k+1}$ there exist matrices $\bar{C}, \bar{\Delta}, C, \Delta$ such that

$$\begin{bmatrix} \Lambda_1 \\ \Lambda_2 \\ \vdots \\ \Lambda_k \end{bmatrix} = \Omega_k \bar{C}', \quad \sigma H_k = \Omega_k \Delta \tag{6.28}$$

and

$$\begin{bmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_k \end{bmatrix} = C \bar{\Omega}_k', \quad \sigma H_k = \Delta \bar{\Omega}_k' \tag{6.29}$$

It is obvious from the last two equalities on the right that there must exist a *unique* matrix A of dimension $n \times n$ such that

$$\sigma H_k = \Omega_k A \bar{\Omega}_k'.$$

In conclusion, the matrices

$$A = \Omega_k^{-L} \sigma H_k (\bar{\Omega}_k')^{-R} \tag{6.30}$$

$$C = \begin{bmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_k \end{bmatrix} (\bar{\Omega}_k')^{-R} \tag{6.31}$$

$$\bar{C} = \begin{bmatrix} \Lambda_1' & \Lambda_2' & \cdots & \Lambda_k' \end{bmatrix} (\Omega_k')^{-R} \tag{6.32}$$

are independent of the choice of the left- or right-inverses (denoted $^{-L}$ or $^{-R}$ respectively) and propagate the factorization (6.27) uniquely to H_{k+1} and $\bar{H}_{k+1}$ according to the formulas,

$$H_{k+1} = \begin{bmatrix} \Omega_k \bar{C}' & \Omega_k A \bar{\Omega}_k' \end{bmatrix} = \Omega_k \begin{bmatrix} \bar{C}' & A \bar{\Omega}_k' \end{bmatrix} := \Omega_k \, \bar{\Omega}_{k+1} \tag{6.33}$$

and

$$\bar{H}_{k+1} = \begin{bmatrix} C \bar{\Omega}_k \\ \Omega_k A \bar{\Omega}_k' \end{bmatrix} = \begin{bmatrix} C \\ \Omega_k A \end{bmatrix} \bar{\Omega}_k' := \Omega_{k+1} \, \bar{\Omega}_k'. \tag{6.34}$$

From these we obtain the following updating equations for the factors Ω_{k+1}, $\bar{\Omega}_{k+1}$,

$$\Omega_{k+1} = \begin{bmatrix} C \\ \Omega_k A \end{bmatrix}, \quad \bar{\Omega}_{k+1} = \begin{bmatrix} \bar{C} \\ \bar{\Omega}_k A' \end{bmatrix}. \tag{6.35}$$

Now once (6.28, 6.29) hold for some $(A, C, \bar{C})$ and k big enough, they must hold with the same $(A, C, \bar{C})$ *for all* $k = 1, \ldots$ and then (6.35) can be interpreted as bona-fide recursions in k. From this we obtain precisely the classical structure of the observabililty and reconstructability matrices

$$\Omega_k = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{k-1} \end{bmatrix} \quad \bar{\Omega}_k = \begin{bmatrix} \bar{C} \\ \bar{C}A' \\ \vdots \\ \bar{C}(A')^{k-1} \end{bmatrix}', \tag{6.36}$$

seen in the literature.

It is important to note that under the equal ranks assumption (6.25), to each rank factorization (6.27) there corresponds a *unique triplet* $(A, C, \bar{C})$. In a sense fixing a rank factorization fixes the basis in the (deterministic) state space of the partial realization. Actually we may amplify this statement in the following way.

Theorem 6.2. *Each rank factorization (6.27) of the finite Hankel matrix H_k satisfying the equal ranks assumption (6.25), determines a unique partial realization $(A, C, \bar{C}, \Lambda(0))$ of the corresponding covariance sequence. Under Assumption 6.1 this realization defines a positive-real function $Z(z)$ and hence each factorization (6.27) determines also a unique uniformly ordered family of stationary realizations of the process $\mathbf{y}$, having the (same) $(A, C, \bar{C})$ parameters given in (6.30, 6.31, 6.32).*

The uniformly ordered family of minimal realizations of y corresponding to a given positive-real quadruple $(A, C, \bar{C}, \Lambda(0))$ was discussed in detail in sections 4. and 4.1.

The "subspace" identification procedure also produces uniformly ordered families of stationary realizations by choosing bases in the finite-memory predictor spaces.

6.5 Partial realization via SVD

One particularly convenient choice of the rank factorization of the Hankel matrix, suggested in [76], later popularized in [44] and refined in [15, 16] is a normalized *Singular-Value factorization.*

Let L_k^- and L_k^+ be the lower triangular Cholesky factors of the block Toeplitz matrices

$$T_k- := \mathrm{E}\{\mathbf{Y}_k^-(\mathbf{Y}_k^-)'\} = L_k^-(L_k^-)' \qquad T_k^+ := \mathrm{E}\{\mathbf{Y}_k^+(\mathbf{Y}_k^+)'\} = L_k^+(L_k^+)'$$

and let

$$\mathbf{e}_k^- := (L_k^-)^{-1}\mathbf{Y}_k^- \qquad \bar{\mathbf{e}}_k^+ := (L_k^+)^{-1}\mathbf{Y}_k^+ \tag{6.37}$$

be the corresponding orthonormal bases in $\mathbf{Y}_k^-$, $\mathbf{Y}_k^+$ respectively (i.e. the finite interval forward and backward innovation vectors).

Introduce the *normalized Hankel matrix*:

$$\hat{H}_k := (L_k^+)^{-1}H_k(L_k^-)^{-T} = \mathrm{E}\,\bar{\mathbf{e}}_k^+(\mathbf{e}_k^-)'$$

and consider the Singular-Value decomposition (SVD) of $\hat{H}_k$,

$$\hat{H}_k = \hat{U}_k \hat{\Sigma}_k \hat{V}_k' \tag{6.38}$$

where $\hat{U}_k$, $\hat{V}_k$ are $mk \times mk$ orthogonal matrices and $\hat{\Sigma}_k$ is diagonal with nonnegative elements[17]

$$1 \geq \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{mk} \geq 0$$

We now do a numerical "rank determination" step which consists in setting equal to zero the canonical correlation coefficients which are smaller than a predetermined "noise treshold level". In this way we substitute for $\hat{\Sigma}_k$ a diagonal matrix of rank n,

$$\hat{\Sigma}_k \simeq \begin{bmatrix} \Sigma_k & 0 \\ 0 & 0 \end{bmatrix}$$

where

$$\Sigma_k = \mathrm{diag}\{\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n\}$$

where the σ_k's are significantly non-zero, and write (with some misuse of noation)

$$\hat{H}_k = U_k \Sigma_k V_k' \tag{6.39}$$

where now U_k and V_k are $mk \times n$ with orthonormal columns.

It is well-known that the "truncated" matrix on the right-hand side of (6.39) provides a best approximation of rank n of H_k in a variety of matrix norms [32]. It is however to be stressed also that this approximation is *no longer Hankel*, or if we prefer the euphemism, only "approximately Hankel".

Since the application of a rigorous Hankel approximation theory [1] would lead to complications, this difficulty is ignored in the following. An analysis of the additional errors involved in this approximation seems still to be an open problem.

From (6.39) a rank factorization of H_k is naturally,

$$\Omega_k := L_k^+ U_k \Sigma_k^{1/2}, \qquad \bar{\Omega}_k := L_k^- V_k \Sigma_k^{1/2}$$

which produces

$$A = \Sigma_k^{-1/2} U_k' (L_k^+)^{-1} \sigma H_k (L_k^-)^{-T} V_k \Sigma_k^{-1/2} \tag{6.40}$$

$$C = \begin{bmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_k \end{bmatrix} (L_k^-)^{-T} V_k \Sigma_k^{-1/2} \tag{6.41}$$

$$\bar{C} = \begin{bmatrix} \Lambda_1' & \Lambda_2' & \cdots & \Lambda_k' \end{bmatrix} (L_k^+)^{-T} U_k \Sigma_k^{-1/2} \tag{6.42}$$

These formulas provide a partial realization of the sequence (6.21) enjoying special properties. Before turning to the analysis of these properties we remark that it may be desirable to rewrite them in a way which is more

[17] These are the well-known sample *canonical correlation coefficients* of the two random vectors $\mathbf{Y}_k^+$ and $\mathbf{Y}_k^-$. In geometric terms they are the cosines of the the *principal angles* between the subspaces $\mathbf{Y}_k^-$ and $\mathbf{Y}_k^+$.

convenient from the numerical point of view, where the explicit computation of the sample covariance matrices is not needed. The SVD computations can be done directly on a suitable QR-type factorization of the Hankel matrices representing the data. A number of other improvements can be introduced for problems of high dimension making (6.40)-(6.42) a quite reliable and fast computational scheme.

6.6 Stochastic Balanced Realizations: the stationary setting

We shall momentarily return to the abstract probabilistic setting of sections 2 and 3. Consider a stationary random process defined on the whole time axis $\mathbb{Z}$, with a rational spectral density $\Phi(z)$ represented as in (3.7). The following definition has been introduced by Desai and Pal in [14].

Definition 6.1. *A minimal realization* $(A, C, \bar{C}, \Lambda(0))$ *of a* $m \times m$ *positive real matrix is called* Stochastically Balanced[18] *if the minimal solutions* P_-, $\bar{P}_+$ *of the dual Linear Matrix Inequalities (4.1), (4.4) are both equal to the same diagonal matrix, i.e.*

$$P_- = \Sigma = \bar{P}_+$$

where $\Sigma = \mathrm{diag}\{\sigma_1, \sigma_2, \dots, \sigma_n\}$. *Whitout loss of generality we shall assume that the* σ_k*'s are ordered in decreasing magnitude, i.e.* $\sigma_{k+1} \geq \sigma_k$.

The motivation of this definition looks rather obscure at this stage and it is not really clear what balanced realizations should be good for. Below we shall provide an explanation based on [67].

Consider a minimal realization of y of the form (2.1) with (minimal) state space X. We shall start our discussion by attaching to each random variable ξ in X a pair of indices which quantify "how well" ξ can be estimated on the basis of the past or future history of the output process. We shall then define a choice of basis in X which has some "canonical" desirable properties in this respect. Initially our discussion will be completely coordinate-free.

For a random variable $\xi \in X$ we define the numbers,

$$\eta_+(\xi) := \frac{\|\mathrm{E}^{H^+}\xi\|^2}{\|\xi\|^2} \qquad \eta_-(\xi) := \frac{\|\mathrm{E}^{H^-}\xi\|^2}{\|\xi\|^2} \tag{6.43}$$

called the *future-* and, respectively, the *past- relative efficiency* of ξ. The numbers $\eta_\pm(\xi)$ are nonnegative and ≤ 1 and in the statistical literature are commonly referred to as the "percentage of explained variance" (of the random variable being estimated). Clearly, the larger $\eta_\pm(\xi)$, the better (in the sense of smaller estimation error variance) will be the corresponding estimate $\mathrm{E}^{H^\pm}\xi$.

[18] Or *Positive-Real Balanced*.

The relative efficiency indices have also a direct system theoretic interpretation in terms of the *observability* and *constructibility* operators associated to X [49] [50], defined respectively as

$$\mathcal{O} : X \to H^+, \qquad \mathcal{O}\xi := \mathrm{E}^{H^+}\xi \tag{6.44}$$

$$\mathcal{C} : X \to H^-, \qquad \mathcal{C}\xi := \mathrm{E}^{H^-}\xi \tag{6.45}$$

In terms of $\mathcal{O}$ and $\mathcal{C}$ the indices $\eta_+(\xi)$ and $\eta_-(\xi)$ may be interpreted as relative "degree of observability" or as relative "degree of constructibility" of $\xi \in X$.

Recall that the observability ad constructibility operators, introduced in geometric realization theory [49] play a somewhat similar role to the observability and reachability operators in deterministic systems theory to charaterize minimality of a state space. In fact the splitting property of a subspace X can be shown to be equivalent to a factorization of the *Hankel operator* of the process y,

$$\mathbb{H} := \mathrm{E}^{H^-}|_{H^+} : H^+ \to H^-$$

through the space X, as [49]

$$\mathbb{H} = \mathcal{C}\mathcal{O}^* \tag{6.46}$$

a fundamental characterization of minimality being that X is a minimal splitting subspace if and only if the factorization (6.46) is canonical, i.e. $\mathcal{C}$ and $\mathcal{O}$ are both *injective* operators. Equivalently (in the finite dimensional case) $\mathcal{O}^* = \mathrm{E}^X|_{H^+}$ is *surjective*. Hence, for a minimal splitting subspace, both $\mathcal{C}^*\mathcal{C}$ and $\mathcal{O}^*\mathcal{O}$ are invertible maps $X \to X$.

It follows that in a minimal splitting subspace X there are two distinct orthonormal bases of eigenvectors, say $(\xi_1^+, \dots \xi_n^+)$ and $(\xi_1^-, \dots \xi_n^-)$ in which the operators $\mathcal{O}^*\mathcal{O}$ and $\mathcal{C}^*\mathcal{C}$ diagonalize, i.e.

$$\mathcal{O}^*\mathcal{O} = \mathrm{diag}\{\lambda_1^+ \dots \lambda_n^+\}, \qquad 1 \geq \lambda_1^+ \geq \dots \geq \lambda_n^+ > 0 \tag{6.47}$$

$$\mathcal{C}^*\mathcal{C} = \mathrm{diag}\{\lambda_1^- \dots \lambda_n^-\}, \qquad 1 \geq \lambda_1^- \geq \dots \geq \lambda_n^- > 0 \tag{6.48}$$

the statistical interpretation being that the states in X can be *ordered* in two different ways according to the magnitude of their future- and, respectively, past- relative efficiency indices. It is in fact immediate from the definition (6.43) that, in the ordering according to the index η_+ the "most observable" states are just those which lie parallel to the vector ξ_1^+, having maximal index $\eta_+(\xi) = \lambda_1^+$ while the "least observable" states ξ being those parallel to the direction ξ_n^+, having the smallest possible relative efficiency $\eta_+(\xi) = \lambda_n^+$. Of course a completely similar picture corresponds to the ordering induced by past–relative efficiency.

Assume for a moment that $H^+ \cap H^- = 0$ (which will be the case if, say, the spectrum of the process is coercive [50]). Then a direction of "very observable" states in X, being at a small relative angle with the future subspace H^+, will generally form a "large" angle with the past subspace H^- and hence give

rise to projections onto H^- of small relative norm i.e. to small $\eta_-(\xi)$. The opposite phenomenon is of course to be expected in case a direction "very close" to H^- is selected. The idea of balancing in the stochastic framework has to do with a choice of basis which roughly speaking, is meant to "balance" i.e. to make equal (if possible), the two ordered sets of efficiency indices. This is meant to reduce as far as possible bad conditioning of the model in the same sense as in determinstic balancing. There is here a substantial difference with the deterministic case however, in that we have now *a whole family of minimal* X which need to be considered simultaneously for the choice of a balanced basis. For this reason the stochastic procedure will necessarily be somehow less obvious than in the deterministic case.

In order to analyze the effects of choosing a basis $x(0)$ in a minimal splitting subspace X, we shall introduce the linear map $T_{x(0)} : \mathbb{R}^n \to X$, defined by $T_{x(0)}a := a'x(0)$. Note that if $\mathbb{R}^n$ is equipped with the inner product $< a, b >_P := a'Pb$, where P is the covariance matrix of $x(0)$, then $T_{x(0)}$ becomes an isometry. From this observation it is not hard to check that $T_{x(0)}$ has the following properties,

Lemma 6.2. *Let P be the covariance matrix of the basis $x(0)$ in X. Then,*

$$T_{x(0)}^{-1} = P^{-1}T_{x(0)}^* \qquad\qquad T_{\bar{x}(0)} = T_{x(0)}P^{-1} \tag{6.49}$$

where $\bar{x}(0)$ is the dual basis of $x(0)$.

Obviously the efficiency indices (6.43) can be expressed in terms of the coordinates a, b, once a specific basis has been chosen. In particular the expressions of the numerators will be quadratic forms described by certain symmetric positive–definite matrices which we shall call, respectively, *Observability* and *Constructibility gramians* (relative to that particular basis). Provided they are expressed in dual bases, the two gramians, have a particularly simple expression that will be given in the Proposition below. Recall (Proposition 4.6) that a basis in an arbitrary X can be extended together with its dual, to the whole family of minimal splitting subspaces $\mathfrak{X}$ in such a way as to form a *uniform basis*.

Proposition 6.1. *Let $x(0)$ be a basis in the minimal splitting subspace X and $\bar{x}(0)$ be its dual basis. Then the constructibility and observability gramians relative to the bases $x(0)$ and $\bar{x}(0)$ respectively, are given by*

$$\hat{\mathcal{C}^*\mathcal{C}} := T_{x(0)}^*\mathcal{C}^*\mathcal{C}T_{x(0)} = P_- \tag{6.50}$$

$$\hat{\mathcal{O}^*\mathcal{O}} := T_{\bar{x}(0)}^*\mathcal{O}^*\mathcal{O}T_{\bar{x}(0)} = \bar{P}_+ = P_+^{-1} \tag{6.51}$$

where P_- and $\bar{P}_+$ are the covariance matrices of $x_-(0)$ and $\bar{x}_+(0)$ in the uniform basis induced by $x(0)$.

In particular the two gramians are invariant over $\mathfrak{X}$, i.e. do not depend on the particular minimal splitting subspace X.

Proof. The formulas follow from the orthogonality of any minimal splitting subspace to the so called "junk" spaces, N^-, N^+ (the subspace of H^- orthogonal to the future and, respectively, the subspace of H^+ orthogonal to the past), see e.g. [50], Corollary 4.9. This leads to the identities

$$\mathcal{C}\xi := \mathrm{E}^{H^-}\xi = \mathrm{E}^{X_-}\xi, \qquad \mathcal{O}\,\xi = \mathrm{E}^{H^+}\xi = \mathrm{E}^{X_+}\xi \tag{6.52}$$

the first of which, in force of (4.11), can be rewritten as $\mathcal{C}\,\xi = a'x_-(0)$ and immediately leads to (6.50). The second follows by a similar computation, using the dual invariant projection property (4.12).

Note that in the forward basis induced by $x(0)$, the expression of the observability gramian would instead be

$$\hat{\mathcal{O}^*\mathcal{O}} := T^*_{x(0)}\mathcal{O}^*\mathcal{O}T_{x(0)} = P\,P_+^{-1}\,P \tag{6.53}$$

which is no longer invariant.

The invariance of the two Gramians with respect to the particular state space of the realization, pointed out in the proposition above, clarifies that the notion of balanced realization given by Desai and Pal in terms of covariance matrices turns (luckily) out to be the correct generalization of the deterministic idea to stochastic systems.

Theorem 6.3. *There is a choice of basis $\hat{x}(0) := [\hat{x}_1, \dots, \hat{x}_n]'$ in X, such that both the constructibility and observability gramians are represented by a diagonal matrix. In fact, there is a diagonal matrix Σ, with positive entries*

$$\Sigma = \mathrm{diag}\{\sigma_1 \dots \sigma_n\}, \qquad 1 \geq \sigma_1 \geq \dots \geq \sigma_n > 0 \tag{6.54}$$

such that, in the uniform basis induced by $\hat{x}(0)$ in $\mathcal{X}$, one has

$$\hat{\mathcal{C}^*\mathcal{C}} = \Sigma = \hat{\mathcal{O}^*\mathcal{O}}, \tag{6.55}$$

where $\hat{\mathcal{C}^*\mathcal{C}}$ is the constructibility gramian relative to the basis $\hat{x}(0)$ and $\hat{\mathcal{O}^*\mathcal{O}}$ is the observability gramian relative to the dual *basis of $\hat{x}(0)$.*

If the numbers σ_k are all distinct, this choice of basis is unique modulo sign, i.e. for any other basis $\tilde{x}(0) := [\tilde{x}_1, \dots \tilde{x}_n]'$ leading to a diagonal structure of the form (6.55), one has $\tilde{x}_k = \pm\hat{x}_k, k = 1, ..., n$.

Recall, as observed in subsection 4.1, that choosing bases uniformly in the family of minimal state spaces $\mathcal{X}$ is equivalent to fixing a (minimal) realization $(A, C, \bar{C}, \Lambda(0))$ of the spectral density matrix. It follows that balanced realizations are generically *canonical forms* with respect to system similarity for (deterministic) realizations $(A, C, \bar{C}, \Lambda(0))$ of the spectrum.

Corollary 6.1. *There always exists a similarity transformation which brings a minimal (positive-real) quadruple $(A, C, \bar{C}, \Lambda(0))$ into balanced form. If the numbers $\{\sigma_1, \sigma_2, \dots, \sigma_n\}$ are all distinct then the balanced realization is unique up to a signature matrix (i.e. any two balanced realizations differ by a change of basis given by a signature matrix).*

For a much deeper discussion of balanced canonical forms see [58, 59].

Algorithm for computing the change of basis matrix bringing a minimal positive realization $(A, C, \bar{C}, \Lambda(0))$ *into balanced canonical form.*

1. Compute a square factorization of P_-, i.e. let $P_- = RR^*$ where R is square nonsingular, e.g. a Cholesky factor.
2. Do Singular Value Decomposition of $R^*\bar{P}_+R$, i.e. compute the factorization $R^*\bar{P}_+R = U\Sigma^2U^*$ where U is an orthogonal matrix and Σ^2 is diagonal with positive entries ordered by magnitude in the decreasing sense.
3. Define $T := \Sigma^{1/2}U^*R^{-1}$. The matrix T is the desired basis transformation matrix.
4. Check: Compute

$$T\,P_-\,T^* = \Sigma^{1/2}U^*R^{-1}\,P_-\,R^{-*}U\Sigma^{1/2} = \Sigma$$

$$T^{-*}\bar{P}_+T^{-1} = \Sigma^{-1/2}U^*R^*\,\bar{P}_+\,RU\Sigma^{-1/2} = \Sigma$$

The meaning of the diagonal matrix Σ. Note that in force of (6.50),(6.51) and (6.55), the numbers $\{\sigma_1^2, \dots, \sigma_n^2\}$ can be computed directly as the the (ordered) eigenvalues of the ratio $P_-\,P_+^{-1}$.

The following statement, which brings up the meaning of the entries of Σ as the (nonzero) singular values of the Hankel operator of the process y, will be reported here for completeness. It has been known for a long time [15], [63]. The proof in the present setup is particularly simple.

Proposition 6.2. *The entries of* $\Sigma = \mathrm{diag}\{\sigma_1, \dots, \sigma_n\}$ *are invariants of the process* y*, equal to the nonzero singular values of the Hankel operator* $\mathbb{H}$*. They coincide with the canonical correlation coefficients of the past and future spaces* H^-, H^+ *of the process* y*.*

Proof. One just needs to notice that $\{\sigma_1^2, \dots, \sigma_n^2\}$ are the eigenvalues of the operator $\mathcal{C}^*\mathcal{C}\mathcal{O}^*\mathcal{O}$, since

$$P_-\,P_+^{-1} = T_{x(0)}^*\mathcal{C}^*\mathcal{C}T_{x(0)}T_{\bar{x}(0)}^*\mathcal{O}^*\mathcal{O}T_{\bar{x}(0)} = T_{x(0)}^*\mathcal{C}^*\mathcal{C}\mathcal{O}^*\mathcal{O}T_{\bar{x}(0)} \tag{6.56}$$

and by (6.49) $T_{\bar{x}(0)}T_{x(0)}^* = I$. On the other hand, the square of the nonzero singular values of $\mathbb{H}$ are the nonzero eigenvalues of $\mathbb{H}^*\mathbb{H}$, and it follows from the factorization (6.46) that the non–zero eigenvalues of $\mathbb{H}^*\mathbb{H}$ are indeed equal to those of $\mathcal{C}^*\mathcal{C}\mathcal{O}^*\mathcal{O}$.

That the singular values of the Hankel operator $\mathbb{H}$ coincide with the canonical correlation coefficients of the process is also quite standard. A formal verification can be found in [55].

In conclusion, in this section we have shown that the concept of stochastic balancing can be seen as a natural generalization of the deterministic idea of balancing of *stable systems*. In the geometric setting however the "stability" (which is necessary for deterministic balancing) of the model does not enter at all, as the choice of a particular state vector $x(0)$ has nothing to do with

the choice of a particular stability (causality) structure of the corresponding realization. The particular causality structure of the model influences instead the computation of the "balancing" basis transformation $\hat{x}(0) = Tx(0)$ of Theorem 6.3. This aspect is discussed in [67].

6.7 Stochastic Balanced Realizations: the case of finite data

The theory presented above only refers to the stationary case. The concept of balancing which applies to identification, where data are always finite, is that of *finite time balancing*.

Definition 6.2. *A minimal realization* $(A, C, \bar{C}, \Lambda(0))$ *of a* $m \times m$ *positive real matrix is called* Stochastically Balanced at time k, *if the solutions* $P_-(k)$, $\bar{P}_+(k)$ *of the dual Riccati equations (5.4), (5.13) started respectively at* $t = 0$ *with initial condition* $\Pi(0) = 0$, *and at time* T *with* $\bar{\Pi}(T) = 0$, *are both equal to the same diagonal matrix, i.e.*

$$P_-(k) = \Sigma(k) = \bar{P}_+(k)$$

where $\Sigma(k) = \mathrm{diag}\{\sigma_1(k), \sigma_2(k), \dots, \sigma_n(k)\}$. *Whitout loss of generality we assume that the* $\sigma(k)$*'s are ordered in decreasing magnitude, i.e.* $\sigma_{i+1}(k) \geq \sigma_i(k)$.

The normalized SVD factorization (6.39) of a *finite* Hankel matrix leads precisely to a finite-time balanced realization.

Proposition 6.3. *The triple (6.40) is stochastically balanced* at time k.

The proof will result from the discussion presented at the end of the next section.

7. The "Subspace Methods" Identification algorithm of Van Overschee and DeMoor.

The general idea of the so-called "subspace methods" for identification of stochastic systems [60], is to operate directly on vector spaces generated by the data.

The system-theoretical background which explains the procedure in (isomorphic) probabilistic terms is exposed in section 5., see in particular Theorem 5.2. The procedure proposed in [60] consists of a number of steps which conceptually can be described as follows.

Given the past and future data spaces $\mathcal{Y}_k^-$, $\mathcal{Y}_k^+$,

1. Form the sample finite-memory predictor spaces $\hat{X}_{k-} = \mathrm{E}^{\mathcal{Y}_k^-}\mathcal{Y}_k^+$ and $\hat{X}_{k+} = \mathrm{E}^{\mathcal{Y}_k^+}\mathcal{Y}_k^-$.

2. Pick *coherent* bases $\hat{\mathbf{x}}(k), \hat{\bar{\mathbf{x}}}(k)$ in $\hat{X}_{k-}$ and $\hat{X}_{k+}$. These bases will define the state at time k of two finite-interval Kalman filter realizations of $\mathbf{y}$.
3. Repeat step n.2 to get coherent bases $\check{\mathbf{x}}(k+1), \check{\bar{\mathbf{x}}}(k-1)$for $\hat{X}_{(k+1)-}$ and $\hat{X}_{(k-1)+}$.
4. Multiply the bases computed in step 3. by a suitable transformation matrix so as they will correspond to $\hat{\mathbf{x}}(k), \hat{\bar{\mathbf{x}}}(k)$ conditionally shifted by one time step (see section 5. for the definition of the conditional shift).
5. Estimate the matrices $(A, C, \bar{C})$ by formulas of the type (2.12, 2.15). These formulas hold for the finite-interval Kalman filter realizations also, see Theorem 7.2 below.

To compute a stationary state-space model, say a forward stationary innovation model (A, C, B_-, D_-), starting from a realization of the spectrum $(A, C, \bar{C}, \Lambda_0)$, the following additional steps are needed.

6. Check $(A, C, \bar{C}, \Lambda_0)$ for positivity. If positivity is not satisfied one may try to re-run the algorithm by varying k and/or n.
7. If $(A, C, \bar{C}, \Lambda_0)$ is positive solve the Algebraic Riccati equation $\Lambda(P) = 0$ and find the unique stabilizing positive-definite solution P_-.
8. Compute (B_-, D_-) by the formulas

$$D_- = (\Lambda_0 - CP_-C')^{1/2}, \quad B_- = (\bar{C}' - AP_-C')(\Lambda_0 - CP_-C')^{-1/2}. \tag{7.1}$$

For pedagogical reasons we have chosen to follow closely the line of thought of [60] albeit, as argued in [55] this procedure involves some redundant computations which can be avoided. In the following sections we shall discuss in detail the basic steps listed above and explain the reasons of the redundancy.

The present time k will be assumed large enough troughout.

7.1 Choosing bases in the predictor spaces

We shall show that there is a one-to-one correspondence between full rank factorizations of the Hankel matrix H_k and coherent choice of bases in the finite-memory predictor spaces $\hat{X}_{k-}$ and $\hat{X}_{k+}$. This correspondence relates the geometric approach of "subspace methods" to the partial realization approach discussed in section 6.4.

Theorem 7.1. *Let $\hat{\mathbf{x}}(k)$, $\hat{\bar{\mathbf{x}}}(k)$ be n-dimensional bases for the finite memory predictor spaces $\hat{X}_{k-}$ and $\hat{X}_{k+}$ and let*

$$\Omega_k \hat{\mathbf{x}}(k) := \mathrm{E}[\mathbf{Y}_k^+ | \hat{\mathbf{x}}(k)], \quad \bar{\Omega}_k \hat{\bar{\mathbf{x}}}(k) := \mathrm{E}[\bar{\mathbf{Y}}_k^- | \hat{\bar{\mathbf{x}}}(k)]. \tag{7.2}$$

Then H_k has the corresponding rank factorizations

$$H_k = \Omega_k \bar{\Delta}_k' = \Delta_k \bar{\Omega}_k'$$

for some $mk \times n$ matrices $\bar{\Delta}_k$, Δ_k with linearly independent columns. If $\hat{\mathbf{x}}(k)$, $\hat{\bar{\mathbf{x}}}(k)$ are a coherent *pair, then $\bar{\Delta}_k = \bar{\Omega}_k$ and $\Delta_k = \Omega_k$.*

Conversely, for each rank factorization (6.27) of the finite Hankel matrix H_k, the n-vectors

$$\hat{\mathbf{x}}(k) = \bar{\Omega}_k{}'(T_k^-)^{-1}\bar{\mathbf{Y}}_k^- \tag{7.3}$$

$$\hat{\bar{\mathbf{x}}}(k) = \Omega_k'(T_k^+)^{-1}\mathbf{Y}_k^+ \tag{7.4}$$

are coherent bases for the finite-memory forward and backward predictor spaces $\hat{X}_{k-}$ and $\hat{X}_{k+}$ respectively.

Proof. That the factorizations of H_k follow from (7.2) is a consequence of the splitting property (5.18) at time k of $\hat{X}_{k-}$ and $\hat{X}_{k+}$. In particular,

$$\mathcal{Y}_k^+ \perp \mathcal{Y}_k^- \,|\, \hat{X}_{k-}$$

which can be rewritten as,

$$\mathrm{E}\mathbf{y}(t)\mathbf{y}(s)' = \mathrm{E}\{\mathrm{E}[\mathbf{y}(t)|\hat{\mathbf{x}}(k)]\,\mathrm{E}[\mathbf{y}(s)|\hat{\mathbf{x}}(k)]'\}$$

for $t = k, \dots, 2k-1$ and $s = k-1, \dots, 0$. This relation arranged in matrix form is the same as $H_k = \Omega_k P(k)\tilde{\Delta}_k'$ where $P(k) := \mathrm{E}\hat{\mathbf{x}}(k)\hat{\mathbf{x}}(k)' > 0$ and $\tilde{\Delta}_k\hat{\mathbf{x}}(k) = \mathrm{E}[\bar{\mathbf{Y}}_k^-|\hat{\mathbf{x}}(k)]$. Letting $\bar{\Delta}_k := \tilde{\Delta}_k P(k)$ yields the first factorization of H_k. The fact that Ω_k and $\bar{\Delta}_k$ are full rank is implied by observability and constructibility (i.e. minimality) of $\hat{X}_{k-}$, since $\hat{\mathbf{x}}(k)$ is a basis. Naturally an analogous reasonig yields the other factorization.

Let $\hat{\mathbf{x}}_+(k) := \bar{P}(k)^{-1}\hat{\bar{\mathbf{x}}}(k)$ be the dual basis of $\hat{\bar{\mathbf{x}}}(k)$ and assume that $\mathrm{E}[\hat{\mathbf{x}}_+(k)|\hat{\mathbf{x}}(k)] = \hat{\mathbf{x}}(k)$ (Proposition 5.2). Since the components of $\hat{\mathbf{x}}_+(k)$ belong to the future we have $\bar{\mathbf{Y}}_k^- \perp \hat{\mathbf{x}}_+(k)\,|\,\hat{\mathbf{x}}(k)$ so that

$$\begin{aligned}\tilde{\Delta}_k P(k) &= \mathrm{E}\{\mathrm{E}[\bar{\mathbf{Y}}_k^-|\hat{\mathbf{x}}(k)]\hat{\mathbf{x}}(k)'\} = \mathrm{E}\{\mathrm{E}[\bar{\mathbf{Y}}_k^-|\hat{\mathbf{x}}(k)]\mathrm{E}[\hat{\mathbf{x}}_+(k)|\hat{\mathbf{x}}(k)]'\}\\ &= \mathrm{E}\{\bar{\mathbf{Y}}_k^-\hat{\mathbf{x}}_+(k)'\} = \mathrm{E}\{\bar{\mathbf{Y}}_k^-\hat{\bar{\mathbf{x}}}(k)'\}\bar{P}(k)^{-1} = \bar{\Omega}_k.\end{aligned}$$

To show the converse, take any random variable in $\mathcal{Y}_k^+$, i.e. any linear combination of the form $a'\mathbf{Y}_k^+, a \in \mathbb{R}^{mk}$ and project it onto $\mathcal{Y}_k^-$. Expressing the projection in terms of the generators $\bar{\mathbf{Y}}_k^-$ of $\mathcal{Y}_k^-$, we obtain

$$\mathrm{E}^{\mathcal{Y}_k^-}\, a'\mathbf{Y}_k^+ = a'H_k(T_k^-)^{-1}\bar{\mathbf{Y}}_k^- = a'\Omega_k\hat{\mathbf{x}}(k)$$

and since the columns of Ω_k are linearly independent it follows that the minimal splitting subspace $\hat{X}_{k-}$ is spanned by the scalar components of $\hat{\mathbf{x}}(k)$. These are also linearly independent since $\hat{\mathbf{x}}(k)$ has a positive definite variance matrix. A dual reasoning for $\hat{\bar{\mathbf{x}}}(k)$ leads to the same conclusion.

That the two bases are coherent is shown in the proposition below.

The variance matrices $P(k) := \mathrm{E}\hat{\mathbf{x}}(k)\hat{\mathbf{x}}(k)'$ and $\bar{P}(k) := \mathrm{E}\hat{\bar{\mathbf{x}}}(k)\hat{\bar{x}}(k)'$ are given by

$$P(k) = \bar{\Omega_k}'(T_k^-)^{-1}\bar{\Omega}_k, \quad \bar{P}(k) = \Omega_k'(T_k^+)^{-1}\Omega_k$$

For future use we record also the formula,

$$\mathrm{E}\hat{\bar{\mathbf{x}}}(k)\hat{\mathbf{x}}(k)' = \bar{P}(k)P(k). \tag{7.5}$$

Proposition 7.1. *The two bases (7.3) and (7.4) are coherent in the sense explained in section 5. i.e. belong to the same uniform choice of bases, or, which is the same, to the same triple* $(A, C, \bar{C})$.

Proof. We shall interpret (with some foresight) $\hat{\bar{\mathbf{x}}}(k)$ as a *dual basis* in $\hat{X}_{k+}$ and $\hat{\mathbf{x}}(k)$ as a "primal basis" in $\hat{X}_{k-}$, the "primal" and dual corresponding bases being,

$$\hat{\mathbf{x}}_+(k) := \bar{P}(k)^{-1}\hat{\bar{\mathbf{x}}}(k), \quad \hat{\bar{\mathbf{x}}}_-(k) := P(k)^{-1}\hat{\mathbf{x}}(k)$$

respectively. Using (7.5) we compute

$$\mathrm{E}^{\hat{X}_{k-}}\hat{\mathbf{x}}_+(k) = \mathrm{E}\{\hat{\mathbf{x}}_+(k)\hat{\mathbf{x}}(k)'\}P(k)^{-1}\hat{\mathbf{x}}(k) = \bar{P}(k)^{-1}\bar{P}(k)P(k)P(k)^{-1}\hat{\mathbf{x}}(k) = \hat{\mathbf{x}}(k)$$

which is the projection condition in Proposition 5.2. So $\hat{\mathbf{x}}(k)$ and $\hat{\bar{\mathbf{x}}}(k)$ are coherent.

Recall that in order to define the same triplet $(A, C, \bar{C})$, the two state vectors (7.3) and (7.4) must "match" i.e. be coherent in the sense explained in section 5..

As we have just seen, a choice of basis (state) vectors in the two spaces $\hat{X}_{k-}$ and $\hat{X}_{k+}$ is related in a one-to-one way to rank-factorizations of the Hankel matrix H_k. Note that by stationarity $\mathbf{y}$ admits also stationary realizations of dimension n of the standard structure (2.1) (see Theorem 5.1) and hence its spectrum is represented by some (minimal) triplet $(A, C, \bar{C})$ of degree n. Information about this triplet is encoded in the bases $\hat{\mathbf{x}}(k)$ and $\hat{\bar{\mathbf{x}}}(k)$, see Theorem 5.2.

We shall now describe a (conceptual) procedure to determine the triplet $(A, C, \bar{C})$ corresponding to an arbitrary choice of bases in the finite memory predictor spaces $\hat{X}_{k-}$, $\hat{X}_{k+}$ as operated above.

The basic idea to compute the dynamics, and in particular the A matrix, is to select bases in the "updated" predictor spaces $\hat{X}_{(k+1)-}$ and $\hat{X}_{(k-1)+}$ constructed with one more observation in the past and one more observation in the future, respectively.

Note however that this further basis selection must be done in such a way as to keep (A, C) and $(\bar{A}, \bar{C})$ constant in time. This is the same as *conditional shifting* defined in section 5.2. Once we know how to do this, the computation of $(A, C, \bar{C})$ is easy.

Theorem 7.2. *Let $\hat{\mathbf{x}}(k), \hat{\bar{\mathbf{x}}}(k)$ be coherent bases in $\hat{X}_{k-}$ and $\hat{X}_{k+}$ and let $\hat{\mathbf{x}}(k+1), \hat{\bar{\mathbf{x}}}(k-1)$ be the corresponding conditionally shifted bases in $\hat{X}_{(k+1)-}$ and $\hat{X}_{(k-1)+}$. The corresponding minimal triple $(A, C, \bar{C})$ can be computed by the following formulas,*

$$A = \mathrm{E}\hat{\mathbf{x}}(k+1)\hat{\mathbf{x}}(k)'\hat{P}(k)^{-1} \tag{7.6}$$

$$C = \mathrm{E}\mathbf{y}(k)\hat{\mathbf{x}}(k)'\hat{P}(k)^{-1} \tag{7.7}$$

$$A' = \mathrm{E}\hat{\bar{\mathbf{x}}}(t-1)\hat{\bar{\mathbf{x}}}(k)'\hat{\bar{P}}(k)^{-1} \tag{7.8}$$

$$\bar{C} = \mathrm{E}\mathbf{y}(k-1)\hat{\bar{\mathbf{x}}}(k)'\hat{\bar{P}}(k)^{-1} \tag{7.9}$$

where $\hat{P}(k) = \mathrm{E}\hat{\mathbf{x}}(k)\hat{\mathbf{x}}(k)'$ and $\hat{\bar{P}}(k) = \mathrm{E}\hat{\bar{\mathbf{x}}}(k)\hat{\bar{\mathbf{x}}}(k)'$.

Proof. The formulas follow readily from the finite interval Kalman filter realizations corresponding to $\hat{\mathbf{x}}(k), \hat{\bar{\mathbf{x}}}(k)$. The fact that $\hat{\bar{\mathbf{x}}}(k)$ and $\hat{\mathbf{x}}(k)$ are coherent bases serves precisely the purpose of extracting the parameters $(A', \bar{C})$ from the backward Kalman filter corresponding to $\hat{\bar{\mathbf{x}}}(k)$.

How do we select conditionally shifted bases? It is obvious that the statement of Theorem 7.1 applies as well to *any* block Hankel matrix constructed with the available covariance data and in particular to the "shifted" Hankel matrices H_{k+1} and $\bar{H}_{k+1}$ defined in (6.23) and (6.24). Assume the rank condition (6.25) holds and consider the Hankel factorizations (6.33, 6.34), namely $H_{k+1} = \Omega_k \bar{\Omega}'_{k+1}$, $\bar{H}_{k+1} = \Omega_{k+1}\bar{\Omega}'_k$ induced by the factorization (6.27) at time k. Corresponding to these factorizations introduce the n-dimensional vectors,

$$\hat{\mathbf{x}}(k+1) = \bar{\Omega}'_{k+1}(T^-_{k+1})^{-1}\bar{\mathbf{Y}}^-_{k+1}, \quad \hat{\bar{\mathbf{x}}}(k+1) = \Omega_k(T^+_k)^{-1}\mathbf{U}\mathbf{Y}^+_k \tag{7.10}$$

$$\hat{\mathbf{x}}(k-1) = \bar{\Omega_k}'(T^-_k)^{-1}\mathbf{U}^{-1}\bar{\mathbf{Y}}^-_k, \quad \hat{\bar{\mathbf{x}}}(k-1) = \Omega'_{k+1}(T^+_{k+1})^{-1}\mathbf{U}^{-1}\mathbf{Y}_{[k,2k]}. \tag{7.11}$$

Now it follows from Theorem 7.1 above that (7.10) are basis vectors for the forward predictor space at time $k+1$ with memory $k+1$: $\hat{X}_{(k+1)-} := \mathrm{E}[\mathbf{U}\mathcal{Y}_{[k,2k-1]}|\mathcal{Y}^-_{k+1}]$, and respectively for the backward predictor space at time $k+1$ with memory of lenght k (in the future), defined as the orthogonal projection $\mathrm{E}[\mathcal{Y}^-_{k+1}|\mathbf{U}\mathbf{Y}_{[k,2k-1]}]$. By Theorem5.1 this projection is actually identical to $\mathrm{E}[\mathbf{U}\mathcal{Y}^-_k|\mathbf{U}\mathcal{Y}_{[k,2k-1]}] = \mathbf{U}\mathrm{E}[\mathcal{Y}^-_k|\mathbf{Y}_{[k,2k-1]}] = \mathbf{U}\hat{X}_{k+}$.

Dually (7.11) are basis vectors, respectively, for the forward predictor space $\mathbf{U}^{-1}\hat{X}_{k-}$ and for the backward predictor space with memory $k+1$ in the future, $\hat{X}_{(k-1)+} := \mathrm{E}[\mathcal{Y}^-_k|\mathcal{Y}_{[k,2k]}]$.

Proposition 7.2. *The random vectors $\hat{\mathbf{x}}(k+1)$, $\hat{\bar{\mathbf{x}}}(k-1)$ defined in (7.10) and (7.11) are the conditionally shifted versions of (7.3) one step forward in time and of (7.4) one step backwards in time.*

Proof. Let A and $\bar{C}$ be the $n \times n$ and $n \times m$ matrices in (6.30), (6.32). We proceed to show directly that

$$\begin{aligned} \mathrm{E}[\hat{\mathbf{x}}(k+1)|\,\hat{\mathbf{x}}(k)\,] &= A\hat{\mathbf{x}}(k) && (7.12)\\ \mathrm{E}[\mathbf{y}(k)\hat{\mathbf{x}}(k+1)'] &= \bar{C} && (7.13) \end{aligned}$$

so that by Proposition 5.3 and by Theorem 6.2 our claim will follow. In fact,

$$\begin{aligned} &\mathrm{E}\hat{\mathbf{x}}(k+1)\hat{\mathbf{x}}(k)'\hat{P}(k)^{-1} = \\ &\bar{\Omega}'_{k+1}(T^-_{k+1})^{-1}\mathrm{E}\bar{\mathbf{Y}}^-_{k+1}(\bar{\mathbf{Y}}^-_k)'(T^-_k)^{-1}\bar{\Omega}_k P(k)^{-1} = \\ &\bar{\Omega}'_{k+1}(T^-_{k+1})^{-1}\mathrm{E}\bar{\mathbf{Y}}^-_{k+1}(\bar{\mathbf{Y}}^-_k)'(\bar{\Omega}'_k)^{-R} \end{aligned}$$

where the last equality follows since from the e**x**pression of $P(k)$ given above,

$$\bar{\Omega_k}'(T^-_k)^{-1}\bar{\Omega}_k P(k)^{-1} = I.$$

Now since

$$(T^-_{k+1})^{-1}\begin{bmatrix} \Lambda_0 & \Lambda_1 & \dots & \Lambda_k \\ \Lambda'_1 & \Lambda_0 & \dots & \Lambda_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda'_k & \Lambda'_{k-1} & \dots & \Lambda_o \end{bmatrix} = \begin{bmatrix} I_m & 0 \\ 0 & I_{(k-1)m} \end{bmatrix}$$

we get,

$$(T^-_{k+1})^{-1}\mathrm{E}\bar{\mathbf{Y}}^-_{k+1}(\bar{\mathbf{Y}}^-_k)' = \begin{bmatrix} 0 \\ I_{(k-1)m} \end{bmatrix}$$

and finally,

$$\begin{aligned} &\bar{\Omega}'_{k+1}\begin{bmatrix} 0 \\ I_{(k-1)m} \end{bmatrix}(\bar{\Omega}'_k)^{-R} = \Omega_k^{-L}H_{k+1}\begin{bmatrix} 0 \\ I_{(k-1)m} \end{bmatrix}(\bar{\Omega}'_k)^{-R} = \\ &\Omega_k^{-L}\sigma H_k(\bar{\Omega}'_k)^{-R} = A. \end{aligned}$$

The verification of (7.13) can be done along similar lines.

From the proof we obtain the following interesting statement,

Corollary 7.1. *The Kalman-Filter realizations having as state vectors the bases (7.3) and (7.4) in the finite-memory predictor spaces $\hat{X}_{k-}$ and $\hat{X}_{k+}$, have the same $A, C, \bar{C}$ parameters as those computed by the partial realization formulas (6.30, 6.31, 6.32) corresponding to the rank-factorization $H_k = \Omega_k\bar{\Omega_k}'$ induced by (7.3) and (7.4) in the sense described in Theorem 7.1.*

In other words, the "subspace methods" algorithm described at the beginning of this section is *equivalent to partial realization*, for the formulas (7.6, 7.7, 7.9) produce exactly the same $(A, C, \bar{C})$ matrices as the partial realization formulas (6.30, 6.31, 6.32) applied to the corresponding Hankel factorization.

Note that the conditionally shifted bases $\hat{\mathbf{x}}(k+1)$ and $\hat{\bar{\mathbf{x}}}(k-1)$ can be computed from the sole factorization (6.27) since $\bar{\Omega}_{k+1}$ and Ω_k are uniquely determined from (6.27) as

$$\bar{\Omega}'_{k+1} = \Omega_k^{-L} H_{k+1}, \quad \Omega_{k+1} = \bar{H}_{k+1}(\bar{\Omega}'_k)^{-R}$$

so that

$$\hat{\mathbf{x}}(k+1) = \Omega_k^{-L} H_{k+1}(T_{k+1}^-)^{-1}\bar{\mathbf{Y}}_{k+1}^- \tag{7.14}$$

$$\hat{\bar{\mathbf{x}}}(k-1) = (\bar{\Omega}_k)^{-L}\bar{H}'_{k+1}(T_{k+1}^+)^{-1}\mathbf{U}^{-1}\mathbf{Y}_{[k,2k]}. \tag{7.15}$$

Change of basis. If we pick arbitrarily an n-dimensional basis $\mathbf{s}(k+1)$ in $\hat{X}_{(k+1)-}$ the basis transformation matrix M taking $\mathbf{s}(k+1)$ into the conditionally shifted basis at time $k+1$ can be obtained by the following reasoning.

First notice that the first members of both expressions

$$\begin{aligned} \mathrm{E}[\mathbf{U}\mathbf{Y}_k^+|\hat{\mathbf{x}}(k+1)\,] &= \Omega_k\hat{\mathbf{x}}(k+1) \\ \mathrm{E}[\mathbf{U}\mathbf{Y}_k^+|\mathbf{s}(k+1)\,] &:= \tilde{\Omega}_k\mathbf{s}(k+1), \end{aligned}$$

are equal to $\mathrm{E}[\mathbf{U}\mathbf{Y}_k^+|\bar{\mathbf{Y}}_{k+1}^-\,]$ by the splitting property. Obviously they must be equal so that $\Omega_k\hat{\mathbf{x}}(k+1) = \tilde{\Omega}_k\mathbf{s}(k+1)$ and

$$\hat{\mathbf{x}}(k+1) = (\Omega_k)^{-L}\tilde{\Omega}_k\mathbf{s}(k+1). \tag{7.16}$$

which provides the change of basis formula in $\hat{X}_{(k+1)-}$. A similar formula can be derived easily for the change of basis in the backward predictor space.

7.2 Skipping some redundant steps

As we have already warned the reader, the procedure for computing $(A, C, \bar{C})$ given so far is vastly redundant from a computational point of view. In principle we can eliminate the computation of the backward bases completely and reduce everything just to finding a basis $\hat{\mathbf{x}}(k)$ in $\hat{X}_{k-}$.

Also there is no need to pick a basis at time $k+1$ in $\hat{X}_{k+1}$ and to convert it to the conditionally shifted basis of $\hat{\mathbf{x}}(k)$, since the conditionally shifted basis $\hat{\mathbf{x}}(k+1)$ can be computed explicitely via formula (7.14). For, choosing a basis $\hat{\mathbf{x}}(k)$ induces a rank factorization (6.27) where the matrix Ω_k is determined by $\hat{\mathbf{x}}(k)$ as shown in (7.2) of Theorem 7.1 above.

Reduced "subspace" Algorithm.

1. Choose a basis $\hat{\mathbf{x}}(k)$ in $\hat{X}_{k-}$.
2. Compute the corresponding observability matrix Ω_k by (7.2).
3. Solve $H_k = \Omega_k \bar{\Omega}_k{}'$ to get (a unique) $\bar{\Omega}_k$.
4. Compute the conditionally shifted basis $\hat{\mathbf{x}}(k+1)$ by (7.14).
5. Compute $(A, C, \bar{C})$ by the following formulas,

$$A = \mathrm{E}\hat{\mathbf{x}}(k+1)\hat{\mathbf{x}}(k)'\hat{P}(k)^{-1} \tag{7.17}$$
$$C = \mathrm{E}\mathbf{y}(k)\hat{\mathbf{x}}(k)'\hat{P}(k)^{-1} \tag{7.18}$$
$$\bar{C} = \mathrm{E}\mathbf{y}(k-1)\hat{\mathbf{x}}(k)' \tag{7.19}$$

where $\hat{P}(k) = \mathrm{E}\hat{\mathbf{x}}(k)\hat{\mathbf{x}}(k)' = \bar{\Omega}_k{}'(T_k^-)^{-1}\bar{\Omega}_k$

Note that (7.19), which formally is derived from the backward (or anti-causal) form of the Kalman Filter realization with state $\hat{\mathbf{x}}(k)$, can be rewritten directly in terms of the dual basis $\hat{\bar{\mathbf{x}}}_-(k) = \hat{P}(k)^{-1}\hat{\mathbf{x}}(k)$ whereby,

$$\mathbf{y}(k-1) = \bar{C}\hat{\bar{\mathbf{x}}}_-(k) + \bar{D}_-(k)\bar{\epsilon}_-(k-1).$$

This reduced procedure should lead to a more effective numerical algorithm than the variants of the original subspace algorithm of [60] which have recently appeared in the literature.

7.3 The least squares implementation

If the "expectation" operator E is written explicitely as in (6.18), then the formulas for $(A, C, \bar{C})$ of Theorem 7.2 express exactly the solution of the two dual *least squares problems,*

$$\min_{A,C} \| \begin{bmatrix} \hat{\mathbf{x}}(k+1) \\ \mathbf{y}(k) \end{bmatrix} - \begin{bmatrix} A \\ C \end{bmatrix} \hat{\mathbf{x}}(k) \|^2 \tag{7.20}$$

$$\min_{A',\bar{C}} \| \begin{bmatrix} \hat{\bar{\mathbf{x}}}(k-1) \\ \mathbf{y}(k-1) \end{bmatrix} - \begin{bmatrix} A' \\ \bar{C} \end{bmatrix} \hat{\bar{\mathbf{x}}}(k) \|^2 \tag{7.21}$$

where the norm is now ordinary Euclidean norm in $\mathbb{R}^N$. This equivalence can be used in the actual computation of $(A, C, \bar{C})$ requiring just a least-squares equation solver. Good numerical implementations for lest-squares problems are easily available. However we should notice that in this formulation we need to compute e**x**plicitely *all* the basis vectors $\hat{\mathbf{x}}(k), \hat{\bar{\mathbf{x}}}(k), \hat{\mathbf{x}}(k+1), \hat{\bar{\mathbf{x}}}(k-1)$.

This rephrasing of the formulas of Theorem 7.2 is used in commercially available codes. The appearence of least squares looks appealing to many and there have been attempts to use the reformulation above also for theoretical purposes. In this respect, there seems to be some confusion in the literature regarding the role played by the estimation residues of the least-squares solution, say

$$\begin{bmatrix} \hat{\mathbf{x}}(k+1) \\ \mathbf{y}(k) \end{bmatrix} - \begin{bmatrix} A \\ C \end{bmatrix} \hat{\mathbf{x}}(k) := \begin{bmatrix} \hat{\mathbf{e}}_{\mathbf{x}}(k) \\ \hat{\mathbf{e}}(k) \end{bmatrix}$$

in "proving" positive-realness of the estimated triple $(A, C, \bar{C})$.

Although it is easy to check that

$$\mathrm{E}\begin{bmatrix} \hat{\mathbf{e}}_{\mathbf{x}}(k) \\ \hat{\mathbf{e}}(k) \end{bmatrix} [\hat{\mathbf{e}}_{\mathbf{x}}(k)'\hat{\mathbf{e}}(k)'] = \begin{bmatrix} P(k+1) - AP(k)A' & \bar{C}' - AP(k)C' \\ \bar{C} - CP(k)A' & \Lambda(0) - CP(k)C' \end{bmatrix} \geq 0$$

there is obviously no guarantee that some $P \geq 0$ will satisfy the stationary matrix inequality $M(P) \geq 0$. To draw this conclusion from the previous expression requires existence of a positive limit of $P(k)$ as $k \to \infty$ which, as is well known, is equivalent to assuming positivity of $(A, C, \bar{C})$ from the beginning.

7.4 Use of the SVD

Of course determinig rank and "picking bases" in practice is a numerically nontrivial affair. The basic numerical tool which helps in this respect is the SVD. In particular the truncated SVD derived from (6.39) of the previous section leads to the choice

$$\Omega_k = L_k^+ U_k \Sigma_k^{1/2}, \quad \bar{\Omega}_k = L_k^- V_k \Sigma_k^{1/2} \tag{7.22}$$

These expressions are meant to be substituted for $\Omega_k, \bar{\Omega}_k$ everywhere in the formulas above in this section wehnever the purpose is to do actual computations.

For the sake of clarity of exposition, in this section we shall assume that the factorization (6.39) is *exact* i.e. that Σ_k is made of the n nonzero singular values of $\hat{H}_k$. From this particular choice of the factorization, the n-dimensional bases for the finite-memory forward and backward predictor spaces $\hat{X}_{k-}$ and $\hat{X}_{k+}$ are seen to be

$$\mathbf{z}(k) = \Sigma_k^{1/2} V_k' (L_k^-)^{-1} \mathbf{Y}_k^- \tag{7.23}$$

$$\bar{\mathbf{z}}(k) = \Sigma_k^{1/2} U_k' (L_k^+)^{-1} \mathbf{Y}_k^+ \tag{7.24}$$

We note immediately that in this basis the variance matrices $P(k)$ and $\bar{P}(k)$ are equal and diagonal,

$$\mathrm{E}\mathbf{z}(k)\mathbf{z}(k)' = \Sigma_k = \mathrm{E}\bar{\mathbf{z}}(k)\bar{\mathbf{z}}(k)'. \tag{7.25}$$

In fact we shall see shortly that $\mathbf{z}(k)$ and $\bar{\mathbf{z}}(k)$ are a (finite-time) *balanced basis.*

Moreover, since $U_k' U_k = I_n = V_k' V_k$ by orthonormality of the columns of U_k and V_k we see that the bases are diagonally correlated i.e.

$$\mathrm{E}\bar{\mathbf{z}}(k)\mathbf{z}(k)' = \Sigma_k^{1/2} U_k' \hat{H}_k V_k \Sigma_k^{1/2} = \Sigma_k^2. \tag{7.26}$$

Hence the vectors $\mathbf{z}(k)$ and $\bar{\mathbf{z}}(k)$ are essentially the so-called *canonical variates* of canonical correlation analysis [37]). The elements of Σ_k are the sample *canonical correlation coefficients* of the finite past and future spaces $\mathcal{Y}_k^-$, $\mathcal{Y}_k^+$. Their dimension n, i.e. the dimension of the predictor spaces, is in reality determined numerically (or statistically) in the truncation step leading to (6.39), by discarding the canonical correlation coefficients which are smaller than a certain "significance level". So the statement about n-dimensional realizability of $\mathbf{y}$ is really "approximate".

Proof of Proposition 6.3. According to the standard notations used in the stationary setting $\hat{\mathbf{x}}(k) = \hat{\mathbf{x}}_-(k)$ and $\hat{\bar{\mathbf{x}}}(k) = \hat{\bar{\mathbf{x}}}_+(k)$ .For any choice of bases $\hat{\mathbf{x}}(k), \hat{\bar{\mathbf{x}}}(k)$ it follows from the Kalman-Filter representations that the relative variance matrices

$$P(k) = \bar{\Omega_k}'(T_k^-)^{-1}\bar{\Omega_k} = P_-(k), \quad \bar{P}(k) = \Omega_k'(T_k^+)^{-1}\Omega_k = \bar{P}_+(k)$$

are the solutions of the (Finite-interval) Riccati equations (5.4), (5.13). From (7.25) we see that $P(k) = \Sigma_k = \bar{P}(k)$, so, as announced in Proposition 6.3, $\mathbf{z}(k), \bar{\mathbf{z}}(k)$ define a *balanced realization* at time k.

It should be stressed however that the formulas of Theorem 7.2 for $\mathbf{z}(k), \bar{\mathbf{z}}(k)$, analog to (6.40), *will not* yield the stationary balanced triple $(A, C, \bar{C})$ described in section 6.6. For getting the system matrices in this form we first need to solve the steady state Algebraic Riccati Equation obtained with the estimated coefficients $(A, C, \bar{C}, \Lambda(0))$, compute the maximal and minimal solutions P_-, P_+ and then apply to $(A, C, \bar{C})$ the balancing algorithm seen in section 6.6.

Acknowledgments

Discussions of G. Picci with colleagues (M. Deistler, J. van Schuppen, J.C. Willems) which took place at the Como NATO-ASI school are gratefully acknowledged.

References

1. V. M. Adamjan, D. Z. Arov and M. G. Krein, Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagi problem, *Math. USSR Sbornik* **15** (1971), 31–73.
2. H. Akaike, *Markovian representation of stochastic processes by canonical variables*, SIAM J. Control **13** (1975), 162–173.
3. H. Akaike, Canonical Correlation Analysis of Time Series and the use of an Information Criterion, in *System Identification, Advances and case studies*, R.K. Mehra and D.L. Lainiotis eds. Academic Press, 1976.
4. N. I. Akhiezer,I. M. Glazman, *Theory of Linear Operators in Hilbert Space*, Ungar, 1966.
5. B. D. O. Anderson, The inverse problem of stationary covariance generation, *J. Statistical Physics* 1:133–147, 1969.

6. B. D. O. Anderson, A System Theory Criterion for Positive–Real Matrices, *SIAM Journal on Control*, 5, 2:171–182, 1967.
7. T. W. Anderson, *Introduction to Multivariate Statistical Analysis*, John Wiley, 1958.
8. K.S. Arun and S.Y. Kung, Balanced approximation of stochastic systems, *SIAM Journal on Matrix Analysis and Applications*, **11**: 42–68, 1990.
9. M. Aoki, *State Space Modeling of Time Series*, 2nd edition, Springer-Verlag, 1991.
10. C. I. Byrnes and A. Lindquist, The stability and instability of partial realizations, *Systems and Control Letters*, **2** (1982), 2301–2312.
11. C. I. Byrnes and A. Lindquist, On the partial stochasic realization problem, to appear.
12. P. E. Caines, *Linear Stochastic Systems*, Wiley, 1988.
13. H. Cramer, *Mathematical Methods of Statistics*, Princeton, 1949.
14. U.B. Desai and D. Pal, A realization approach to stochastic model reduction and balanced stochastic realization *Proc 16th Annulal Conference on on Information Sciences and Sytems*, Princeton Univ, pp. 613–620, 1982, also in *Proc 21st Conference on Decision and Control*, Orlando, FL, pp.1105–1112, 1982.
15. U.B. Desai and D. Pal, A realization approach to stochastic model reduction *IEEE Transactions Automatic Control*,**AC-29**: 1097–1100, 1984.
16. U.B. Desai, D. Pal and R.D. Kikpatrick, A realization approach to stochastic model reduction *International Journal of Control*, 42: 821–838 1985.
17. U. B. Desai, *Modeling and Application of Stochastic Processes*, Kluwer Academic Publishers, 1986.
18. J.L. Doob, The Elementary Gaussian Processes *Annals of Math. Statistics*, 15: 229–282, 1944.
19. J.L. Doob, *Stochastic Processes*, Wiley, 1953.
20. M. P. Ekstrom, A spectral characterization of the ill-conditioning in numerical deconvolution, *IEEE Trans, Audio Electroacustics*, **AU-21**, pp. 344-348, 1973.
21. P. Faurre, *Identification par minimisation d'une representation Markovienne de processus aleatoires*, Symposium on Optimization, Nice 1969.
22. P. Faurre, P. Chataigner, Identification en temp reel et en temp differee par factorisation de matrices de Hankel, *Proc. French-Swedish colloquium on process control*, IRIA Roquencourt, 1971.
23. P. Faurre and J. P. Marmorat, Un algorithme de réalisation stochastique, *C. R. Academie Sciences Paris* , **268** (1969).
24. P. Faurre, *Representation Markovienne des processus stochastiques stationnaires*, INRIA Report de recherche, 1973
25. P. Faurre, M. Clerget, F. Germain, *Opérateurs Rationnels Positifs*, Dunod, 1979.
26. F. R. Gantmacher, *Matrix Theory*, Vol. I, Chelsea, New York, 1959.
27. K.F. Gauss, *Theoria Motus Corporum Coelestium*, Liber II, in *Werke*, Julius Springer, Berlin, 1901.
28. T. T. Georgiou, *Realization of power spectra from partial covariance sequences*, IEEE Transactions Acoustics, Speech and Signal Processing **ASSP-35** (1987), 438–449.
29. M.R. Gevers and B.D.O. Anderson Representation of jointly stationary feedback free processes, *Intern. Journal of Control* **33**, (1981), pp.777-809.
30. M.R. Gevers and B.D.O. Anderson On jointly stationary feedback free stochastic processes, *IEEE Trans. Automatic Control* **AC-27**, (1982), pp.431-436.
31. K. Glover, All optimal Hankel norm approximations of linear multivariable systems and their L^∞ error bounds. *International Journal of Control*, **39**, 6:1115–1193, 1984.

32. G. H. Golub and C. R. Van Loan, *Matrix Compuations* (2nd ed.). The Johns Hopkins Univ. Press (1989).
33. C.W.J. Granger, Economic processes involving feedback, *Information and Control* **6**, (1963), pp. 28-48.
34. M. Green, Balanced stochastic realizations *Linear Algebra and its Applications*, 98:211–247, 1988.
35. B. R. Hunt, A theorem on the difficulty of numerical deconvolution, *IEEE Trans, Audio Electroacustics*, **AU-20**, March 1972.
36. Ch. Heij, T. Kloek and A. Lucas, *Positivity conditions for stochastic state space modelling of time series*, Reprint Series 695, Erasmus University Rotterdam.
37. H. Hotelling, Relations between two sets of variables, *Biometrica*, **28** (1936), pp. 321–377.
38. P. Harshavaradhana, E. A. Jonckheere and L. M. Silverman, Stochastic balancing and approximation-stability and minimality, *IEEE Trans. Automatic Control*, **AC-29** (1984), 744–746.
39. P. Harshavadana and E.A. Jonckheere Spectral factor reduction by phase–matching, the continuous–time case. *International Journal of Control*, 42: 43–63, 1985.
40. P. Opdenacker and E.A. Jonckheere, A state space approach to to approximation by phase–matching in *Modelling, Identification and Robust Control* (C. I. Byrnes and A. Lindquist eds), Elsevier, 1986.
41. R. E. Kalman, Realization of covariance sequences, *Proc. Toeplitz Memorial Conference*, Tel Aviv, Israel, 1981.
42. R.E.Kalman, P.L.Falb, and M.A.Arbib, *Topics in Mathematical Systems Theory*, McGraw-Hill, 1969.
43. H. Kimura, Positive partial realization of covariance sequences, *Modelling, Identification and Robust Control* (C. I. Byrnes and A. Lindquist, eds.), North-Holland, 1987, pp. 499–513.
44. S. Y. Kung, A new identification and model reduction algoritm via singular value decomposition, *Proc. 12th Asilomar Conf. Circuit, Systems and Computers*, 1978, pp. 705–714.
45. W. E. Larimore, System identification, reduced-order filtering and modeling via canonical variate analysis, *Proc. American Control Conference*, 1990, pp. 445–451.
46. W.E. Larimore, Canonical Variate Analysis in Identification, Filtering, and Adaptive Control. *Proc. 29th IEEE Conference on Decision and Control* (1990), pp. 596–604.
47. A. Lindquist, G. Picci and G. Ruckebusch On minimal splitting subspaces and Markovian representation, *Math. System Theory*, **12**: 271-279, 1979.
48. A. Lindquist and G. Picci, On the stochastic realization problem *SIAM J. Control and Optimization*, **17**: 365–389, 1979.
49. A. Lindquist and G. Picci, Realization theory for multivariate stationary Gaussian processes, *SIAM J. Control and Optimization*, **23**:809–857, 1985.
50. A. Lindquist and G. Picci, A geometric approach to modelling and estimation of linear stochastic systems, *Journal of Mathematical Systems, Estimation and Control*, **1**:241–333, 1991.
51. A. Lindquist, G. Michaletzky and G. Picci, Zeros of Spectral Factors, the Geometry of Splitting Subspaces, and the Algebraic Riccati Inequality, *SIAM J. Control & Optimization* (March 1995).
52. A. Lindquist and G. Michaletzky, Output-induced subspaces,invariant directions and interpolation in linear discrete-time stochastic systems, *Tech Report TRITA/MAT-94-20*, Royal Institute of Technology, Stockholm (1994).

53. A. Lindquist and G. Picci, *On "subspace methods" identification,* in Systems and Networks: Mathematical Theory and Applications II, U. Hemke, R. Mennicken and J Saurer, eds., Akademie Verlag, 1994, pp. 315–320.
54. A. Lindquist and G. Picci, *On "subspace methods" identification and stochastic model reduction,* Proceedings 10th IFAC Symposium on System Identification, Copenhagen, June 1994, Volume 2, pp. 397–403.
55. A. Lindquist and G. Picci, Canonical Correlation Analysis Approximate Covariance Extension and Identification of Stationary Time Series, *Tech Report TRITA/MAT-94-32*, Royal Institute of Technology, Stockholm. (submitted to *Automatica*).
56. B. P. Molinari, The time-invariant linear-quadratic optimal-control problem, *Automatica*, 13:347–357, 1977.
57. B.P.Molinari, The stabilizing solution of the discrete algebraic Riccati equation, *IEEE Trans. Automatic Control*, **20** (1975), 396–399.
58. R. Ober, Balanced realizations: canonical forms, parametrization, model reduxction, *International Journal of Control* **46** (1987), pp. 643-670.
59. R. Ober, Balanced parametrization of a class of linear systems, *SIAM Journal on Control & Optimization*, 29, 6:1251–1287, 1991.
60. P. van Overschee and B. De Moor, *Subspace algorithms for stochastic identification problem*, Automatica **3** (1993), 649-660.
61. P. van Overschee and B. De Moor, *Two subspace algorithms for the identification of combined deterministic-stochastic systems*, preprint.
62. P. van Overschee and B. De Moor, A unifying theorem for subspace identification algorithms and its interpretation, *Proceedings 10th IFAC Symposium on System Identification*, Copenhagen, June 1994, Volume 2, pp. 145–156.
63. M. Pavon, Canonical Correlations of past inputs and future outputs for linear stochastic systems *Systems and Control Letters*, **4**: 209–215, 1984.
64. L. Pernebo and L. M. Silverman, Model reduction via balanced state space representations, *IEEE Trans. Automatic Control*, **AC-27** (1982), 382–387.
65. D. L. Phillips, A technique for the numerical solution of certain integral equations of the first kind, *Journal of the Assoc. Comput. Mach.*, **9** pp. 97-101, 1962.
66. G. Picci, Stochastic realization of Gaussian Processes, *Proceedings of the IEEE*, **64** (1976), pp. 112-122.
67. G. Picci and S. Pinzoni, Acausal models and balanced realizations of stationary processes, *Linear Algebra and its Applications*, **205-206** (1994), 957-1003.
68. N. I. Rozanov, *Stationary Random Processes*, Holden Day, 1963.
69. A.Tether, Construction of minimal state-variable models from input-output data, *IEEE Trans. Automatic Control* **AC-15** (1971), pp. 427-436.
70. S. Twomey, The application of nunerical filtering to the solution of integral equations of the first kind encountered in indirect sensing measurements, *Journal of the Franklin Institute*, **279**, pp. 95-109, 1965.
71. R. J. Vaccaro and T. Vukina, A solution to the positivity problem in the state-space approach to modeling vector-valued time series, *J. Economic Dynamics and Control* **17** (1993), pp. 401–421.
72. S. Weiland, Theory of Approximation and disturbance attenuation for linear systems *Doctoral Thesis*, University of Groningen, Jan 1991.
73. N. Wiener, Generalized Harmonic Analysis, in *The Fourier Integral and Certain of its Applications*, Cambridge U.P. 1933.
74. N. Wiener and P. Masani, The prediction theory of multivariate stationary stochastic processes, I, *Acta Mathematica***98**, 11-150, (1957); II, *ibidem*, **99** 93-137, 1958.

75. J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation,* IEEE Trans. Automatic Control **AC-16** (1971), pp. 621–634.
76. H. P. Zeiger and A. J. McEwen, *Approximate linear realization of given dimension via Ho's algorithm,* IEEE Trans. Automatic Control **AC-19** (1974), p. 153.
77. D.C. Youla, on The Factorization of Rational Matrices, *IRE Transactions PGIT* , **7**: 1961.