1

GEOMETRIC METHODS FOR STATE SPACE IDENTIFICATION

Giorgio Picci

April 22, 2005

[1]Department of Information Engineering, University of Padua, via Gradenigo 6/A, 35131 Padua, Italy

Contents

Preface

This is the preface. This is the preface. This is the preface. This is the preface. This is the preface. This is the preface. This is the preface. This is the preface.

Chapter 1
Introduction

We have nothing to fear but fear itself.
—Franklin D. Roosevelt

I am not a crook.
—Richard M. Nixon

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 observable 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.

1.1 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 don't want to) take into account 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 bunch 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 philosophycal discussion in [27] sect. III, p. 236.

More to the point, there has been a diffused 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 small variations in 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 underlying 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, forcing a solution of the integral equation (1.1) from real data leads normally to disastrous results. This is by now very well-known and documented in the literature, see e.g. [65, 69, 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 bounds on the magnitude of the relative "uncertainty region" around the nominal model. This type of modeling philosophy has been put forward in view of applications to H^∞ control. 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 don't 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 distinction among theories of model building is 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.

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

[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.

[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.

procedure there must be a formalization of the belief that

> *future date 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 does offer a logical standpoint to build a theory for assessing the *quality* of the identified model, by *asymptotic analysis*, i.e. 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, like e.g. stationary data, where a precise justification for the adoption of the stochastic description of uncertainty can be given. In what follows we shall attempt to offer a formal argument to motivate this choice.

First, in order to capture the idea that future data are "statistically the same" as past data we shall introduce a definition of *stationarity* of a (deterministic) signal.

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 any 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 operator* σ is a map defined as

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

transforming a signal z into its "translation by t units of time" $z_t := \{z(t+s)\}_{s\in\mathbb{Z}_+}$. The shift can be made to act also on functions of z according to the rule

$$\sigma^t f(z) := f(z(t+s); s \in I) = f(z_t).$$

In the following we shall denote $\sigma^t f(z)$ by the more compact notation $f_t(z)$.

Definition 1.1. *The 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;

- Second-order *stationary if the limit exists for $f(z) = z(0)$ (so that $f_t(z) = z(t)$) and for all quadratic forms in z, i.e. for all real functions f such that $f(\alpha z) = \alpha^2 f(z)$.*

The definition extends in a natural way to vector-valued sequences.

The two conditions in the definition of a second-order stationary signal represent more or less the minimum amount of structure necessary to do a rudimental asymptotic analysis of an identification algorithm for linear stationary models. They are normally found in the literature under a variety of different names.

We shall now show that (strict-sense) stationary signals admit a natural mathematical description as trajectories of *stationary stochastic processes*.

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 stationary sequences we have $\nu_T(A) \to \nu_0(A)$ as $T \to \infty$. It follows readily from the observation above that

Lemma 1.2. *The set function $A \to \nu_0(A)$ is a probability measure on $\mathbb{R}$.*

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 1.2 leads to the following result.

Lemma 1.3. *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 then 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[3]. In other words,

Proposition 1.4. *For a stationary signal 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 the signal 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.*

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 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*. Moreover we shall assume throughout that the time averages of all signals are subtracted off so

[3]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).

all data will 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 1.5 (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 [72]. □

From this result, much in the same spirit of the strict-sense result above, one can draw the conclusion that a wide-sense stationary signal admits as 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 to form a stationary[4] signal z; we shall call Λ_0 the *true covariance* of the 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.

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 $\Lambda_0(\tau)$ forms a summable sequence (so that $\sum |\Lambda_0(\tau)| < \infty$) then the spectral distribution function admits a *density* Φ_0, and

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

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

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 the previous section, a reasonable mathematical description of the measured data is to model it as a chunck of a trajectory of a second-order wide-sense stationary stochastic process. The identification problem is then naturally formulated as the problem of recovering the "true" second order law of the process i.e. its true covariance or spectral distribution function from the observed trajectory.

Instead, in a *strict-sense* formulation one would try, much more ambtiously, to infer the true probability law of the underlying process from the measured data.

These are of course just prototypical problems of Statistics. In this book we shall only deal with the second-order formulation.

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, it is natural to study identification of *finite dimensional* model classes which are commonly assumed in the prediction and control schemes of the engineering literature. Although we keep the meaning of the term somewhat vague at this stage, it is very well-known that wide-sense stationary processes describable by "finite-dimensional models" can only be linear combinations of quasi-periodic (i.e. sums of a finite number of sinusoids with random amplitudes) and purely-non-deterministic processes with a *rational spectral density*. Generally, after a proper data pre-processing of the observed time series, one may well assume that the observed process $\mathbf{y}$ is a *purely non-deterministic process*. It is well known that this property is equivalent to assuming that for no $a \in \mathbb{R}^n$, $a^\top \mathbf{y}(t)$ can be expressed exactly as a linear combination of components of past variabes $\mathbf{y}(t-1), \mathbf{y}(t-2), \dots$ of the process. From this it can be easily shown that the block Toeplitz matrix

$$T_k := \begin{bmatrix} \Lambda(0) & \Lambda(1) & \Lambda(2) & \cdots & \Lambda(k) \\ \Lambda(1)^\top & \Lambda(0) & \Lambda(1) & \cdots & \Lambda(k-1) \\ \Lambda(2)^\top & \Lambda(1)^\top & \Lambda(0) & \cdots & \Lambda(k-2) \\ \vdots & \vdots & \vdots & & \\ \Lambda(k)^\top & \Lambda(k-1)^\top & \Lambda(k-2)^\top & \cdots & \Lambda(0) \end{bmatrix} \tag{1.4}$$

must be *(strictly) positive definite for all k*. There is then very little choice for the model class. If we are interested in finite-complexity modeling of "truly random" (purely-non-deterministic) signals, then we must restrict to processes admitting a *rational spectral density*. These modeling issues will be discussed in Chapter **??**.

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.

As we shall better see later, in the present second-order stochastic setup, 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 an additive structure of the form

$$y(t) = F(\zeta)u(t) + v(t) \tag{1.5}$$

where $F(\zeta)u(t)$ is, in symbolic notation, a causal and stable linear system (a convolution operator) with a rational transfer function $F(\zeta)$. The additive term $v(t)$ is the "stochastic component", a stationary process, also with a rational spectrum, uncorrelated with the past of u, which models precisely the uncertainty due to disturbances etc. superimposed to the input-based prediction of $y(t)$.

Note that the above model class comes out as a formal consequence of the probabilistic setting used to describe our data. It can in fact be justified by a trivial application of Wiener filtering theory. There is no arbitrariness or "user choice" at this stage, except of course in the choice of the order or the structure parameters of the transfer function. Note incidentally that identifying the model uncertainty in (1.5) means in particular 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 may even be "assumed" that u is a "deterministic" signal. This of course cannot be the real intention since it would lead to the rather absurd consequence 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*.

Uncorrelation is more likely 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 (thereby 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 (1.5), involving now the "output" process y playing in turn the role of an exogenous variable (i.e. an "input") to determine u.

The appropriate setup for discussing these matters is within the theory of *feedback* and *causality* between stationary processess [12]. We shall not adventure into this subject in this introduction. 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 this reason that we shall first choose to restrict the scope of our discussion to time-series identification. Feedback and feedback models will be studied in Chapter **??**.

1.2 Basic issues in subspace identification of time-series

We shall now discuss some general issues of the statistical identification problem of describing an observed m-dimensional time series

$$\{y_0, y_1, y_2, \dots y_N\}, \tag{1.6}$$

by a finite-dimensional model of the type commonly considered in the engineering or econometrics literature.

As we will see in the next chapters there are different choices of the model class which could be used (and are in fact 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 this model the familiar form of a linear difference equation

   $$y(t) + \sum_{k=1}^{\nu} A_k y(t-k) = \sum_{k=0}^{\nu} B_k w(t-k) \tag{1.7}$$

   also called an "ARMA" model. Among several (minimal) shaping filters, or ARMA model representations, we must choose one. Typically a convenient model to choose is the *innovation* model where W is minimum phase and w is the innovation preocess of y. As we shall see at the end of Section **??**, 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, or identifiable, parametrization*, i.e. of making the correspondence: parameter $\rightarrow$ probabilistic model, generically bijective. For example, the ARMA innovations model must be parametrized

in such a way as to yield an *identifiable parametrization*. This means that one should parametrized by the coefficients $\{A_k, B_k\}$ of the matrix polynomials $(A(z^{-1}),\ B(z^{-1}))$. 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 1970's and early 1980's but is now a bit obsolete after *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.

There are basically two different approaches to the problem of fitting a model to the data,

- The *optimization approach*, based on the principle of minimizing a suitable distance function between the data[5] and the probability law corresponding to 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 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.

- The so-called *Subspace identification approach*. This is a two steps procedure which in principle can be described as construction of a state process for the observed process followed by linear regression and then by a noise parameters identification step which requires the solution of a Riccati equation. This method is based on stochastic realization theory.

[5]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.

From a statistical point of view the difference with the first approach is that the estimation of the model parameters is *not* done by optimizing a likelihood or other distance functions but can simply be seen as *matching second order moments*. For instance, let

$$\{\Lambda_0, \Lambda_1, \ldots, \Lambda_\nu\} \tag{1.8}$$

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 (1.6). The problem is of finding a minimal value of n and a minimal[6] triplet of matrices $(A, C, \bar{C})$, of dimensions $n \times n$, $m \times n$ and $m \times n$ respectively, such that

$$CA^{k-1}\bar{C}' = \Lambda_k \quad k = 1, 2, \ldots, \nu \tag{1.9}$$

The solution of these equations can be accomplished by modern versions of the famous Ho-Kalman algorithm which are simple and numerically reliable. Estimation by solving (3.72) is an istance of *estimation by the method of moments* described in the statistical textbooks [13, p. 497], which is a very old idea, for example used extensively by K. Pearson in the beginning of the 20th century. The underlying priciple is close in spirit to the wide-sense setting that we are working in this book. It does not guarantee anything like minimal distance between the "true" and the model distributions but rather imposes that the parameter estimates 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 $N \to \infty$. In other words the first ν lag values of the true covariance function will be matched exactly.

On the other hand 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 optimization 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 roughly amounts to the same) and this in the long run the optimization generally leads to matching covariances anyway. Asa drawback there is instead the structural handycap of iterative optimization methods which may get stuck in local minima and may well provide sub-optimal parameter estimates, a rather hard phenomenon to detect. The subspace approach offers in this respect the major advantage of converting the nonlinear parameter estimation phase which is the core of maximum-likelihood or prediction-error model identification to 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] Recall that $(A, C, \bar{C})$ is minimal if (A, C) is completely observable and $(A, \bar{C}')$ is completely reachable.

There is one point of warning however. In the literature the covariance matching problem is invariably treated as a *minimal partial realization* problem. The triplet $(A, C, \bar{C})$ is computed by minimal factorization of the block Hankel matrix corresponding to the data (1.8) as follows:

$$H = \begin{bmatrix} \Lambda_1 & \Lambda_2 & \Lambda_3 & \cdots & \Lambda_j \\ \Lambda_2 & \Lambda_3 & \Lambda_4 & \cdots & \Lambda_{j+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda_i & \Lambda_{i+1} & \Lambda_{i+2} & \cdots & \Lambda_{i+j-1} \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{i-1} \end{bmatrix} \begin{bmatrix} \bar{C} \\ \bar{C}A' \\ \vdots \\ \bar{C}(A')^{j-1} \end{bmatrix}', \quad (1.10)$$

where $i + j = \nu$ and $|i - j| \leq 1$. An infinite sequence

$$\{\Lambda_0, \Lambda_1, \Lambda_2, \ldots\} \quad (1.11)$$

is then obtained from (1.8) by setting $CA^{k-1}\bar{C}' = \Lambda_k$ for $k = \nu + 1, \nu + 2, \ldots$, this sequence is called a *minimal rational extension* of the finite sequence (1.8). The elements of (3.73) are the coefficients of the Laurent expansion of the rational function

$$Z(z) = C(zI - A)^{-1}\bar{C}' + \frac{1}{2}\Lambda_0 = \frac{1}{2}\Lambda_0 + \Lambda_1 z^{-1} + \Lambda_2 z^{-2} + \ldots \quad (1.12)$$

about $z = \infty$.

Now $Z(z)$ is expected to be the positive-real part of a spectral density matrix, but the usual deterministic realization algorithms based on factorization of a Hankel matrix do not take into account any positivity constraints. In fact the rational function (3.74) obtained by solving the partial realization equations (3.72) may not only fail to be positive-real but the relative A matrix may even fail to be stable [10]. So the second approach introduces some nontrivial mathematical questions related to positivity of the estimated spectrum. Therefore there is a price to be paid for the simplification allowed by the two-steps approach.

Note that positivity is the natural condition insuring solvability of the Linear Matrix Inequality (or, in particular, of the Riccati equation) required to compute state-space models of the signal from the covariance estimates.

The correct approach would in principle require to compute a rational *positive extension* of the finite covariance sequence (1.8), 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 algorithm, these methods produce functions of very high complexity, in fact generically of the highest possible degree. 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 [**?**].

In these notes we shall be content of discussing the deterministic partial realization approach. Once a positive triple $(A, C, \bar{C})$ is estimated, getting a state-space model is just a matter of solving the LMI or the appropriate Riccati equation as seen

in section 2.3. If the method breaks down (typically the Riccati equation fails to have a solution becouse of nonpositive-realness of the estimated $(A, C, \bar{C})$) there is a signal which th emethod gives, pointing to inadequacy of the selected parameters (regression horizons and model order)

Historical remarks Stochastic realization based identification 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] and Verhaegen [**?**]. 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 section **??**, **??** and on the other hand seems completely unrelated to the partial realization and covariance extension approach mentioned above.

In this book we shall analyze the geometric "Subspace" 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.

Chapter 2
State Space Models of Stationary Processes

Wide-sense stationary second-order processes with a *rational spectral density matrix* provide a natural class of finitely-parametrized stochastic models which are useful for the identification of a wide class of observed data. The scope of this chapter is to study these models in some detail.

2.1 State space models

In this and in the following two sections we shall review the basic facts about finite-dimensional state-space models of stationary random processes[7].

Consider a linear stochastic system

$$(\Sigma) \quad \left\{ \begin{array}{lcl} \mathbf{x}(t+1) & = & A\mathbf{x}(t) + B\mathbf{w}(t) \\ \mathbf{y}(t) & = & C\mathbf{x}(t) + D\mathbf{w}(t) \end{array} \right. \tag{2.1}$$

where (A, B, C, D) are constant matrices and $\{\mathbf{w}(t)\}$ is p-dimensional normalized white noise , i.e.

$$\mathbb{E}\{\mathbf{w}(t)\mathbf{w}(s)^\top\} = I\delta_{ts} \quad \mathbb{E}\{\mathbf{w}(t)\} = 0.$$

In this book we shall think of (2.1) exclusively as a *representation* of the output process $\mathbf{y}$. For this reason it will be often called a *stochastic realization* of the process $\mathbf{y}$. This representation involves *auxiliary variables* such as the *state process* $\mathbf{x}$ and the *generating white noise* $\mathbf{w}$ which are processes of a simpler structure than $\mathbf{y}$ and are introduced at the purpose of giving the model a particular structure and particular properties. These auxiliary variables may be chosen in different ways or even eliminated producing a different model structure. For example, by eliminating $\mathbf{x}$ from the equations (2.1) one obtains an "input-output" representation whereby $\mathbf{y}$ appears as the result of processing the white noise signal $\mathbf{w}$ through a linear time-invariant filter

[7]The material in the two next sections is a slightly remastered and compressed version of previous joint work with Anders Lindquist [48][49][50]. Proofs will be skipped whenever available in the original sources.

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

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

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

and

$$\mathbf{y}(t) = \sum_{j=-\infty}^{t-1} CA^{t-1-j}B\mathbf{w}(j) + D\mathbf{w}(t)$$

In particular, $\mathbf{x}$ and $\mathbf{y}$ are *jointly stationary*[8].

The system (2.1) can be regarded as a linear map defining $\mathbf{x}(t)$ and $\mathbf{y}(t)$ as linear functionals of the input noise $\mathbf{w}$. In fact, whenever the matrix A is stable, $\mathbf{x}(t)$ and $\mathbf{y}(t)$ will depend only on the *past* of the input noise, i.e. only on the random variables $\{\mathbf{w}(s);\, s \le t\}$. We shall say that, in this case, the map is a *causal* map.

Consider the linear vector space of second order zero-mean random variables generated by the scalar components of the process $\mathbf{w}$, i.e. all finite linear combinations of the infinite family of (scalar) random variables $\{\mathbf{w}_i(t) \mid t \in \mathbb{Z}; i = 1, 2, \ldots, p\}$. The closure of this vector space with respect to the norm induced by the inner product $\langle \xi, \eta \rangle = \mathbb{E}\{\xi\eta\}$ is an (infinite dimensional) Hilbert space denoted $H(\mathbf{w})$. Convergence in this space is commonly called convergence in *mean square*. We shall write

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

Here the notation $\overline{\text{span}}$ will always denotes the closure in mean square of the vector space generated by linear combinations of the random variables listed inside the brackets. Likewise, the symbol $H(\mathbf{y})$ will denote the Hilbert space generated by an arbitrary wide-sense zero mean process $\mathbf{y}$. It will be convenient to think of the (components of) $\mathbf{x}$ and $\mathbf{y}$ as elements of $H(\mathbf{w})$. Since all random quantities related to the model (2.1) belong to $H(\mathbf{w})$, this space is called the *ambient space* of the stochastic system (Σ).

For any stationary process $\mathbf{y}$, we can formally define a *shift operator* U, a linear map which is initially defined on the random variables of the form $a^\top \mathbf{y}(t)$, as temporal translation i.e. $Ua^\top \mathbf{y}(t) = a^\top \mathbf{y}(t+1)$. The map is then extended

[8] Stationarity here is always meant in the "wide sense" of second order statistics. In particular $\mathbf{x}$ and $\mathbf{y}$ being jointly stationary means that the covariance matrix $\mathbb{E}\{[\mathbf{x}(t)^\top \mathbf{y}(t)^\top]^\top [\mathbf{x}(s)^\top \mathbf{y}(s)^\top]\}$ depends only on $t - s$.

by linearity and continuity to the whole Hilbert space $H(\mathbf{y})$ [67]. Note that U is norm preserving (as the variances of each component of $\mathbf{y}(t)$ and $\mathbf{y}(t+\tau)$ are equal) and the extension is in fact a *unitary operator*, that is, a linear operator which preseves inner product and is (automatically one to one and) onto $H(\mathbf{y})$. The pair $(H(\mathbf{y}), U)$ 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. Of course the ambient Hilbert space $H(\mathbf{w})$ can likewise be equipped with a unitary shift U with respect to which the processes $\mathbf{x}$ and $\mathbf{y}$ are also stationary.

The *past subspaces* at time t of $\mathbf{x}$ and $\mathbf{y}$

$$H_t^-(\mathbf{x}) = \overline{\text{span}}\{\mathbf{x}_i(s) \mid s < t; i = 1, 2, \dots, n\} \tag{2.5}$$

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

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

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

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

The finite dimensional subspace of $H(\mathbf{w})$

$$X_t = \text{span}\{\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)\} \quad t \in \mathbb{Z},$$

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

$$P = \mathbb{E}\{\mathbf{x}(0)\mathbf{x}(0)^\top\},$$

it follows from stationarity that $P = \mathbb{E}\{\mathbf{x}(t)\mathbf{x}(t)^\top\}$ for all t, and hence the first equation in (2.1) yields

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

which is a Lyapunov equation. Since $|\lambda(A)| < 1$ the sum $P = \sum_{j=0}^{\infty} A^j BB^\top (A^\top)^j$, converges and, as it is easy to show, it is a solution of (2.7). In fact, P is just the reachability Gramian of Σ. But (A, B) is reachable, and hence $P > 0$. This implies that $\{\mathbf{x}_1(t), \mathbf{x}_2(t), \dots, \mathbf{x}_n(t)\}$ 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 $\mathbf{y}$. The orthogonal projection onto a subspace S will be denoted $\mathbb{E}^S$ or $\mathbb{E}[. \mid S]$. For example, if $\boldsymbol{\xi} \in H(\mathbf{w})$ and $Z \subset H(\mathbf{w})$ is spanned by the components of the random vector $\mathbf{z}$, then assuming the components of $\mathbf{z}$ are linearly independent

$$E^Z\boldsymbol{\xi} = E\{\boldsymbol{\xi}\mathbf{z}'\}(E\{\mathbf{z}\mathbf{z}'\})^{-1}\mathbf{z}. \tag{2.8}$$

For Gaussian random variables this coincides with the conditional expectation given the σ-algebra generated by S. The orthogonal projection of a subspace A onto another subspace B is

$$\mathbb{E}^B A := \overline{\text{span}}\,\{\mathbb{E}^B a \mid a \in A\}$$

Operators like $\mathbb{E}^S$ or U are also applied to vector valued random variables with the understanding that in this case they will act on the single components in an obvious way.

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

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 fixed stationary Hilbert space H of second-order random variables containing $H(\mathbf{y})$. Define*

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

The subspace X is Markovian *if the variables in the past, X^-, and in the future, X^+, are conditionally uncorrelated (i.e. orthogonal) given X, which is written as*

$$X^- \perp X^+ \mid X. \tag{2.9}$$

A Markovian Splitting Subspace *X for the process $\mathbf{y}$ is a subspace of H making the joint past $H^- \vee X^-$ and the joint future space $H^+ \vee X^+$ conditionally uncorrelated (i.e. orthogonal) given X, denoted,*

$$H^- \vee X^- \perp H^+ \vee X^+ \mid X. \tag{2.10}$$

A Markovian Splitting subspace is of course Markovian. Any basis vector $\mathbf{x}(0) := [\mathbf{x}_1(0), \mathbf{x}_2(0), \ldots, \mathbf{x}_n(0)]^\top$ in a Markovian splitting subspace X generates a

stationary Markov process $\mathbf{x}(t) := U^t\mathbf{x}(0), t \in \mathbb{Z}$ which serves as a *state process* of the the process $\mathbf{y}$ see [**?**] for a deeper discussion of this concept.

A subspace $X \subset H$ is called *proper*, or *purely non deterministic* if there are vector white noise processes $\mathbf{w}$ and $\bar{\mathbf{w}}$ such that

$$X^- = H^-(\mathbf{w}), \qquad X^+ = H^+(\bar{\mathbf{w}})$$

A stationary process $\mathbf{y}$ is similarly called *purely non deterministic.* if H^- and (respectively) H^+ can be represented as the past space (at time zero) and the future space (at time zero) of a vector white noise processes. Clearly, if both X and the process $\mathbf{y}$ are purely non deterministic then

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

where the white noises $\mathbf{w}$ and $\bar{\mathbf{w}}$ are in general different than those previously encountered (since they must now generate larger spaces). In fact we have

$$H^- \vee X^- = H^-(\mathbf{z}) \qquad \text{where} \quad \mathbf{z}(t) := \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t-1) \end{bmatrix} \tag{2.11}$$

We recall from [67], that an equivalent characterization of joint pure-non-determinism is that

$$\cap_t H_t^-(\mathbf{z}) = \{0\}, \quad \text{and} \quad \cap_t H_t^+(\mathbf{z}) = \{0\}.$$

.

The fundamental characterization in this setting is the following.

Theorem 2.2. *The state space X of any stochastic ralization (2.1) is a Markovian Splitting Subspace for the process $\mathbf{y}$.*

Conversely, given a p.n.d. process $\mathbf{y}$ and any proper Markovian splitting subspace X for $\mathbf{y}$, of finite dimension n, to any choice of basis $\mathbf{x} = [\,\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\,]^\top$ in X there corresponds a stochastic ralization of $\mathbf{y}$ of the type (2.1).

Proof. We shall only prove the converse statement. Let $\mathbf{z}$ be the joint vector process defined in (2.11) whose past space is, by assumption, generated by some vector white noise $\mathbf{w}$. Decompose $H_{t+1}^-(\mathbf{w}) = H_t^-(\mathbf{w}) \oplus H(\mathbf{w}(t))$, which by assumption is the same as $H_{t+1}^-(\mathbf{z}) = H_t^-(\mathbf{z}) \oplus H(\mathbf{w}(t))$. This leads to a decomposition of $\mathbf{z}(t+1) \in H_{t+1}^-(\mathbf{z})$ as

$$\begin{bmatrix} \mathbf{x}(t+1) \\ \mathbf{y}(t) \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} \mathbf{x}(t+1) \\ \mathbf{y}(t) \end{bmatrix} \mid H_t^-(\mathbf{z})\right] + \mathbb{E}\left[\begin{bmatrix} \mathbf{x}(t+1) \\ \mathbf{y}(t) \end{bmatrix} \mid \mathbf{w}(t)\right]$$

By the Markovian splitting property, in the first projection we can substitute $H_t^-(\mathbf{z}) = H_t^- \vee X_t^-$ with just the present state $X_t \equiv H(\mathbf{x}(t)$. The linearity of the projections implies that there are matrices A, C and B, D such that

$$\mathbb{E}\left[\begin{bmatrix} \mathbf{x}(t+1) \\ \mathbf{y}(t) \end{bmatrix} \mid X_t\right] = \begin{bmatrix} A \\ C \end{bmatrix} \mathbf{x}(t), \qquad \mathbb{E}\left[\begin{bmatrix} \mathbf{x}(t+1) \\ \mathbf{y}(t) \end{bmatrix} \mid \mathbf{w}(t)\right] = \begin{bmatrix} B \\ D \end{bmatrix} \mathbf{w}(t).$$

which leads to a state-space model of the type (2.1). □

A particular instance of this representation is rephrased in the corollary below. Note that, although we use the same generic symbol $\mathbf{w}$ for white noise, the two white processes in Theorem 2.2 and in the corollary below will in general be different.

Corollary 2.3. *If $\mathbf{x}$ is a basis in a finite-dimensional proper Markovian subspace X, the Markov process $\mathbf{x}(t) := U^t\mathbf{x}$ is purely non determinstic and can be represented by a linear equation of the type*

$$\mathbf{x}(t+1) = A\mathbf{x}(t) + B\mathbf{w}(t)$$

If (A, B) is a reachable pair, then A has all its eigenvalues strictly inside of the unit circle.

Proof. Since $\mathbf{x}$ is a basis, its variance matrix $P := \mathbb{E}\{\mathbf{x}(t)\mathbf{x}(t)^\top\}$ is positive definite and satisfies the Lyapunov equation $P = APA^\top + BB^\top$. By reachability it must hold that A is stable. □

The coefficient matrices A, C are uniquely determined by the choice of basis in the state space while B, D also depend on the choice of the generating noise $\mathbf{w}$. There are simple formulas expressing them in terms of $\mathbf{x}$ and $\mathbf{y}$ given in (2.12) below.

$$A = \mathbb{E}\,\mathbf{x}(t+1)\mathbf{x}(t)^\top P^{-1} \qquad B = \mathbb{E}\,\mathbf{x}(t+1)\mathbf{w}(t)^\top \tag{2.12a}$$

$$C = \mathbb{E}\,\mathbf{y}(t)\mathbf{x}(t)^\top P^{-1} \qquad D = \mathbb{E}\,\mathbf{y}(t)\mathbf{w}(t)^\top. \tag{2.12b}$$

A Markovian splitting subspace is *minimal* if it doesn' t contain (properly) other Markovian splitting subspaces. Contrary 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):

$$X_- := \mathbb{E}^{\,H^-} H^+ \qquad X_+ := \mathbb{E}^{\,H^+} H^- \tag{2.13}$$

for which we have the following characterization [49].

Proposition 2.4. *The subspaces X_- and X_+ are minimal Markovian splitting subspaces. In fact, they are the minimal Markovian splitting subspaces contained in the past H^-, and, respectively, in the future H^+, of the process $\mathbf{y}$.*

The abstract definitions (2.13) are valid for infinite dimensional (nonrational) processes. If the process has a rational spectrum and hence admits a (minimal) realization of dimension n, then the subspaces X_- and X_- are finitely generated. In fact,

Proposition 2.5. *If the process $\mathbf{y}$ admits a realization (2.1) of dimension n then*

$$(X_t)_- = \operatorname{span}\{\mathbb{E}\,\left[\mathbf{y}(t+k) \mid H_t^-\right];\, k = 0, 1, \ldots, \nu\} \tag{2.14}$$

where ν is the observability index[9] *of the pair (A, C). In fact, X_- is generated by the n components of the (one step ahead predictor) estimate of the state of any realization of $\mathbf{y}$ of the type (2.1). In formulas,*

$$(X_t)_- = \text{span}\,\{\hat{\mathbf{x}}_k(t);\, k = 1, 2, \ldots, n\}, \qquad \hat{\mathbf{x}}(t) := \mathbb{E}\left[\mathbf{x}(t) \mid H_t^-\right]$$

The subspace X_+ admits a dual characterization exchanging past with future.

The estimate $\hat{\mathbf{x}}(t) := \mathbb{E}\left[\mathbf{x}(t) \mid H_t^-\right]$ is actually the state process of the *innovation* or *steady-state Kalman filter realization* which will be discussed at the end of this Chapter.

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

$$H_t^+(\mathbf{w}) \perp X_t^- \vee H_t^-(\mathbf{y}) \tag{2.15}$$

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 $\mathbb{E}\,\{\mathbf{x}(t)\mathbf{w}(t)^\top\} = 0$.

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 $\mathbf{y}$ 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 [66].

Theorem 2.6. *[48, 47] Let $\bar{\mathbf{x}}$ be any basis in X and let $\bar{\mathbf{x}}(t) = U^t\bar{\mathbf{x}}$; $t \in \mathbb{Z}$ be the corresponding stationary vector Markov process. The joint process $\begin{bmatrix} \bar{\mathbf{x}}(t) \\ \mathbf{y}(t) \end{bmatrix}$ is also Markov and admits a* backward representation

$$\begin{bmatrix} \bar{\mathbf{x}}(t-1) \\ \mathbf{y}(t-1) \end{bmatrix} = \begin{bmatrix} \bar{A} \\ \bar{C} \end{bmatrix} \bar{\mathbf{x}}(t) + \begin{bmatrix} \bar{B} \\ \bar{D} \end{bmatrix} \bar{\mathbf{w}}(t-1) \tag{2.16}$$

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

$$\bar{A} = \mathbb{E}\,\bar{\mathbf{x}}(t-1)\bar{\mathbf{x}}(t)^\top \bar{P}^{-1} \quad \bar{B} = \mathbb{E}\,\bar{\mathbf{x}}(t)\bar{\mathbf{w}}(t)^\top \tag{2.17}$$

$$\bar{C} = \mathbb{E}\,\mathbf{y}(t-1)\bar{\mathbf{x}}(t)^\top \bar{P}^{-1} \quad \bar{D} = \mathbb{E}\,\mathbf{y}(t)\bar{\mathbf{w}}(t)^\top \tag{2.18}$$

where $\bar{P} = \mathbb{E}\,\bar{\mathbf{x}}(t)\bar{\mathbf{x}}(t)^\top$.

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

$$\mathbb{E}\,\bar{\mathbf{x}}(t)\mathbf{x}(t)^\top = I$$

[9] The smallest integer r for which

$$\text{rank}\begin{bmatrix} C \\ CA \\ \vdots \\ CA^r \end{bmatrix} = n.$$

which implies

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

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

$$\bar{A} = A^\top \quad \bar{C}' = APC^\top + BD^\top \tag{2.19}$$

Proof. . The proof is symmetric to that of Theorem 2.2 and is based on the orthogonal decomposition $H^+_{t-1}(\mathbf{z}) = H^+_t(\mathbf{z}) \oplus H(\bar{\mathbf{w}}(t-1))$ which implies

$$\begin{bmatrix}\bar{\mathbf{x}}(t-1)\\ \mathbf{y}(t-1)\end{bmatrix} = \mathbb{E}\left[\begin{bmatrix}\bar{\mathbf{x}}(t-1)\\ \mathbf{y}(t-1)\end{bmatrix} \mid H^+_t(\mathbf{z})\right] + \mathbb{E}\left[\begin{bmatrix}\bar{\mathbf{x}}(t-1)\\ \mathbf{y}(t-1)\end{bmatrix} \mid \bar{\mathbf{w}}(t-1)\right]$$

By the Markovian splitting property, in the first projection we can substitute $H^+_t(\mathbf{z}) = H^+_t \vee X^+_t$ by the present state $\bar{X}_t \equiv H(\bar{\mathbf{x}}(t))$. The linearity of the projections implies that there are matrices $\bar{A}, \bar{C}$ and $\bar{B}, \bar{D}$ such that

$$\mathbb{E}\left[\begin{bmatrix}\bar{\mathbf{x}}(t-1)\\ \mathbf{y}(t-1)\end{bmatrix} \mid X_t\right] = \begin{bmatrix}\bar{A}\\ \bar{C}\end{bmatrix}\bar{\mathbf{x}}(t), \qquad \mathbb{E}\left[\begin{bmatrix}\bar{\mathbf{x}}(t-1)\\ \mathbf{y}(t-1)\end{bmatrix} \mid \mathbf{w}(t)\right] = \begin{bmatrix}\bar{B}\\ \bar{D}\end{bmatrix}\mathbf{w}(t).$$

which leads to a state-space model of the type (2.16). Formulas (2.19) follow by using the forward model expressions in (2.17). □

The forward and backward realizations are asymmetric because of the asymmetry in the definition of past and future of $\mathbf{y}$. This asymmetry is needed in order to avoid unnecessarily high state space dimension due to overlap of past and future spaces of the process. For, with the symmetric choice of including the present both in the future and in the past, a p-dimensional white noise process would have a minimal realization with a state space of dimension p. The choice here is to have the present only in $H^+(\mathbf{y})$. However the Markov property requires instead symmetry of past and future so, *for Markov processes*, it is convenient and natural to have the present in *both* past and future spaces.

Now the past space at time zero of the joint process

$$\begin{bmatrix}\mathbf{x}(t)\\ \mathbf{y}(t-1)\end{bmatrix}$$

is $X^- \vee H^-$ which checks with the Markov convention for the past (however the future of the joint process does not span $X^+ \vee H^+$ but instead $X^+ \vee H^+_{-1}$). Dually the future space at time zero of

$$\begin{bmatrix}\bar{\mathbf{x}}(t)\\ \mathbf{y}(t)\end{bmatrix}$$

is exactly $X^+ \vee H^+$ according to the Markov convention. However the past space of the barred process again is not chosen according to the Markov convention being equal to $X^- \vee H^-_{-1}$ (which does not include the present of $\mathbf{y}$).

A more symmetric expression for the backward realization is obtained by shifting time forward by one unit and introducing a shifted state $\boldsymbol{\xi}(t) := \bar{\mathbf{x}}(t+1)$ so that (2.16) can be rewritten

$$\begin{bmatrix} \boldsymbol{\xi}(t-1) \\ \mathbf{y}(t) \end{bmatrix} = \begin{bmatrix} \bar{A} \\ \bar{C} \end{bmatrix} \boldsymbol{\xi}(t) + \begin{bmatrix} \bar{B} \\ \bar{D} \end{bmatrix} \bar{\mathbf{w}}(t)$$

These models will be useful to derive the backward Kalman filter later.

2.2 Spectral Factorization

We have so far collected enough evidence to the fact that, even if we restrict to *minimal realizations* 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 $\mathbf{y}$. In fact we may have minimal representations in which the input noise processes have different dimensions. This is a significant departure from the usual deterministic linear modeling setup and brings up a problem of *model choice* which should be well understood before discussing any statistical methodology for identification. Motivated by this observation, in this section we shall study the family of shaping filter representations (2.2) of the process $\mathbf{y}$.

The covariance sequence of a process $\mathbf{y}$ admitting a representation of the form (2.1), i.e.

$$\Lambda(k) := \mathbb{E}\{\mathbf{y}(t+k)\mathbf{y}(t)^\top\} = \mathbb{E}\{\mathbf{y}(t)\mathbf{y}(0)^\top\}$$

is readily computed using the results of the previous section. It is easy to see that

$$\Lambda(k) = CA^{k-1}\bar{C}^\top \quad \text{for } k > 0, \qquad \Lambda(0) = CPC^\top + DD^\top \tag{2.20}$$

where,

$$\bar{C}^\top = APC^\top + BD^\top. \tag{2.21}$$

is exactly the C matrix of the "backward" model (2.19), which checks with the reverse time covariance expression

$$\Lambda(-k) = \Lambda(k)^\top = \bar{C}(A^\top)^{k-1}C^\top \qquad \text{for } k > 0.$$

Hence it follows that the infinite block Hankel matrix

$$H_\infty := \begin{bmatrix} \Lambda(1) & \Lambda(2) & \Lambda(3) & \cdots \\ \Lambda(2) & \Lambda(3) & \Lambda(4) & \cdots \\ \Lambda(3) & \Lambda(4) & \Lambda(5) & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} \tag{2.22}$$

admits a facorization

$$H_\infty := \begin{bmatrix} C \\ CA \\ CA^2 \\ CA^3 \\ \vdots \end{bmatrix} \begin{bmatrix} \bar{C} \\ \bar{C}A^\top \\ \bar{C}(A^\top)^2 \\ \bar{C}(A^\top)^3 \\ \vdots \end{bmatrix}^\top \tag{2.23}$$

and hence has finite rank bounded above by the dimension n of the state space X_t of the system Σ. Whether or not rank $H_\infty = n$ depends on the reachability of the pair $(A, \bar{C}^\top)$, which, as we shalll see later, is equivalent (assuming that (A, C) is observable) to *stochastic minimality* of the realization (2.1) of the process $\mathbf{y}$ [47, 49, 50]. Note that both (A, C) and $(A^\top, \bar{C})$ are observable, if and only if the deterministic realization of Λ in (2.20) is minimal. Hence,

Proposition 2.7. *If both (A, C) and $(A^\top, \bar{C})$ are observable, the backward state-output matrix $\bar{C}$ is uniquely determined by the forward parameters (A, C).*

This clearly follows since, under the stated assumptions, fixing one of the two factors in the factorization (2.23) uniquely determines the other.

For the purely non-deterministic process $\mathbf{y}$, the spectral distribution is absolutely continuous [67] and admits a density. In our case the $m \times m$ *spectral density* of $\mathbf{y}$ can even be computed as an ordinary Fourier (or z-) 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 since $\Lambda(-k) = \Lambda(k)^\top$, $\Phi(z)$ has the property

$$\Phi(1/z) = \Phi(z)^\top$$

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

$$\Phi(z) = \Phi_+(z) + \Phi_+(1/z)^\top \tag{2.24}$$

where $\Phi_+(z)$ is the transform of the "causal" tract (2.20) of the covariance. For symmetry we have assigned one half of the constant term $\Lambda(0)$ to the causal part of Λ and the other half to the anticausal component. It is evident that of $\Phi_+(z)$ is a rational matrix function, analytic outside of the unit circle, given by the expression

$$\begin{aligned} \Phi_+(z) &= \frac{1}{2}\Lambda(0) + \Lambda(1)z^{-1} + \Lambda(2)z^{-2} + \cdots \\ &= C(zI - A)^{-1}\bar{C}^\top + \frac{1}{2}\Lambda(0). \end{aligned} \tag{2.25}$$

The positivity condition of the sequence of Toplitz matrices (1.4) is equivalent to positive semidefiniteness of $\Phi(z)$ on the unit circle i.e.

$$\Phi_+(e^{j\theta}) + \Phi_+(e^{-j\theta})^\top \geq 0 \quad \theta \in [-\pi, \pi] \tag{2.26}$$

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

Proposition 2.8. *The transfer function W of any state space representation of the process $\mathbf{y}$ of the type (2.1) is a* spectral factor *of Φ, i.e.*

$$W(z)W(1/z)^\top = \Phi(z). \tag{2.27}$$

There is a straightforward proof of this result in case of the A matrix is stable, 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 $W(z)W(1/z)^\top$ 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 (2.24) and (2.25) and need not have a probabilistic meaning).

Proof. A straightforward calculation shows that

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

Now, using a famous trick apparently invented by Kalman and Yakubovich, bring in the identity

$$P - APA^\top = (zI-A)P(z^{-1}I-A^\top) + (zI-A)PA^\top + AP(z^{-1}I-A^\top), \quad (2.28)$$

which, in view of (2.7), yields

$$\begin{aligned} W(z)W(1/z)^\top &= CPC^\top + DD^\top + C(zI-A)^{-1}(APC^\top + BD^\top) \\ &+ (CPA^\top + DB^\top)(z^{-1}I-A^\top)^{-1}C^\top \\ &= \Phi_+(z) + \Phi_+(1/z)^\top. \end{aligned} \quad (2.29)$$

where the last equality follows from (2.21). □ □

Note that the proof only requires existence of a solution to the Lyapunov equation $P = APA^\top + BB^\top$. 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 *analyitic* spectral factor.

We shall need to consider also *antistable* spectral factors $\bar{W}(z)$, i.e. (rational) solutions of the spectral factorization equation (2.27), having all poles ouside of the unit circle. These spectral factors can be characterized as the stable factors $G(z)$ of the transpose spectrum $\Phi(z)^\top$, subjected to the transformation of variable

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

so that $\bar{W}(z)\bar{W}(1/z)^\top = G(1/z)G(z)^\top = \Phi(z)$. Antistable spectral factors turn out to be exactly the transfer functions of backward realizations of $\mathbf{y}$, i.e. state-space representations of the form (2.16). For, the transfer function of a backward model (2.16) 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 (2.25) of Φ_+ induces a natural transpose realization for the transpose

$\Phi_+(z)^\top$, namely

$$\Phi_+(z)^\top = \bar{C}(zI - A^\top)^{-1}C^\top + \frac{1}{2}\Lambda(0), \tag{2.30}$$

we see that the dual choice of basis of Theorem 2.6 for the backward models is a natural one. Hence by just switching symbols according to the correspondence

$$A \leftrightarrow A^\top \quad C \leftrightarrow \bar{C},$$

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 reachability and observability of (A, B, C) in (2.1), the pair $(A, \bar{C}^\top)$ *may not be reachable* and hence

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

may not be a minimal realization.

We recall that the *McMillan degree* $\delta(F)$ of a proper rational matrix function $F(z)$ is just the dimension of a minimal realization of $F(z)$. We have the following proposition relating the McMillan degree of a spectral factor to that of Φ. The proof can be found in Anderson's paper [5].

Proposition 2.9. *For any spectral factor W, it holds that*

$$\delta(W) \geq \frac{1}{2}\delta(\Phi) = \delta(\Phi_+). \tag{2.31}$$

Whenever equality holds, we say that W is a *minimal spectral factor* of Φ.

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* (i.e. analytic in $\{|z| < 1\}$ factors with all the zeros outside or, respectively, inside of the unit circle, denoted[10] $\bar{W}_+$ and $\bar{W}_-$ respectively. The factor $\bar{W}_+$ is commonly called *conjugate minimum-phase* or *co-outer*.

Theorem 2.10. *All stable rational spectral factors can be constructed by postmultiplying the minimum phase factor by a stable rational matrix function $Q(z)$ such that*

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

[10]The rationale for the subscripts will become clear 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

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

Transfer function like Q or $\bar{Q}$ are called *all-pass* . Stable and square all-pass matrix functions are called *inner*. The result above goes back to Youla' s classical 1961 paper [76].

2.3 Spectral Factorization and the LMI

In this section we consider the problem of computing all *minimal stable* spectral factors W of a rational spectrum by computing the corresponding (minimal) realizations say $W(z) = D + H(zI - F)^{-1}B$. (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, we shall assume we are given a minimal realization

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

where $J + J^\top = \Lambda(0)$ and A is a stable matrix. We shall solve the spectral factorization equation (**??**), giving a procedure to compute (F, H, B, D) from the "data" $(A, C, \bar{C}, \Lambda(0))$.

From the expression we have found earlier for the covariance function it should be clear that F and H could be chosen equal to A and C for all factors. Hence the problem can be reduced to finding just the B and D matrices. This is the content of the following theorem.

Theorem 2.11. *Let $(A, C, \bar{C}^\top)$ be a minimal realization of the causal part of the spectrum. 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

$$M(P) := \begin{bmatrix} P - APA^\top & \bar{C}^\top - APC^\top \\ \bar{C} - CPA^\top & \Lambda(0) - CPC^\top \end{bmatrix} \geq 0 \tag{2.32}$$

in the following sense:

Corresponding to each solution $P = P^\top$ of (2.32), which is necessarily positive definite, consider the full column rank factorization $M(P)$,

$$M(P) = \begin{bmatrix} B \\ D \end{bmatrix} [B^\top D^\top] \tag{2.33}$$

and the rational matrix W parametrized in the form

$$W(z) = C(zI - A)^{-1}B + D. \tag{2.34}$$

Then (2.34) 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* $P = P^\top$ *of the Lyapunov equation (2.7) satisfy the matrix factorization equation (2.33) and hence the Linear Matrix Inequality (2.32).*

Proof. Let $P = P^\top$ be a solution of (2.32) and B, D be computed as in (2.33). Then P solves the Lyapunov equation (2.7) and hence $P > 0$. Then forming the product $W(z)W(1/z)^\top$ it follows from the equation (2.29) 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 (2.31). 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 $P = P^\top > 0$ exists solving the Lyapunov equation $BB^\top = P - FPF^\top$ and hence from the spectral factorization equation and the Kalman-Yakubovich identity (2.29) we get

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

which implies that Φ_+ is also realized by a necessarily minimal (since we are considering minimal spectral factors for which $\delta(W) = \delta(\Phi_+)$) matrix triple of the form $(F, H, \bar{H})$. Therefore(F, H) and (A, C) are similar and we may take $F = A$ and $H = C$ for all minimal spectral factors.

Rearranging the Lyapunov equation for P, the definition of $\bar{C}$, and the expression $\Lambda(0) - CPC^\top = DD^\top$ in matrix form (see equation (**??**) it is then obvious that (P, B, D) satisfy (2.33) and hence P is a positive definite solution of (2.32). □

Note that the full rank factors in (2.33) are unique only modulo right multiplication by orthogonal transformations, but this is of no harm since the same indeterminacy holds for spectral factors.

The equations (2.33) are sometimes called the *positive real equations*, and can be written

$$\underbrace{\begin{bmatrix} P - APA^\top & \bar{C}^\top - APC^\top \\ \bar{C} - CPA^\top & \Lambda(0) - CPC^\top \end{bmatrix}}_{M(P)} = \begin{bmatrix} B \\ D \end{bmatrix} \begin{bmatrix} B^\top & D^\top \end{bmatrix} \geq 0 \tag{2.35}$$

so we may look at the linear function $M : \mathbb{R}^{n\times n} \to \mathbb{R}^{2n\times 2n}$, which depends on the known parameters $(A, \bar{C}, C, \Lambda(0))$, as a map from P to (B, D) pairs.

One immediate consequence of Theorem 2.11 is that the size of the minimal spectral factors can be computed from the rank of the corresponding matrix $M(P)$. In fact if we agree to consider only solutions $\begin{bmatrix} B \\ D \end{bmatrix}$ of full column rank, it follows

from the the factorization above that the corresponding $W(z)$ must be $m \times p$ with $p = \operatorname{rank} M(P)$.

It can be shown [25] that the set of solutions to the LMI (2.32)

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

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 2.11 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. (2.26) holds.*

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

The Dual Positive-Real Equations A dual of Theorem 2.11 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*

$$\bar{M}(\bar{P}) := \begin{bmatrix} \bar{P} - A^\top \bar{P} A & C^\top - A^\top \bar{P} \bar{C}^\top \\ C - \bar{C}\bar{P}A & \Lambda(0) - \bar{C}\bar{P}\bar{C}^\top \end{bmatrix} \geq 0 \tag{2.36}$$

can readily be obtained by replacing the realization $\Phi_+(z) = C(zI - A)^{-1}\bar{C}^\top + J$, by the transpose realization representing $\Phi_+(z)^\top$ 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 (2.36) there corresponds an antistable minimal spectral factor

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

where $\bar{B}, \bar{D}$ are determined by the analog of the matrix factorization (2.33).

In the following we shall assume that

$$R(P) := \Lambda(0) - CPC^\top > 0 \tag{2.37}$$

for all $P \in \mathcal{P}$. This means that all minimal state space models of $\mathbf{y}$ have a full-rank additive noise term in the output equation. In other words, the D matrix is full rank and $DD^\top > 0$. This condition serves here only the purpose of avoiding the use of pseudo-inverses and of simplifying the exposition. It is curious that a natural characterization of the spectra for which this condition holds has taken a long time to emerge in the literature (see however the recent paper [**?**]). Under this assumption, if $T := -(\bar{C}^\top - APC^\top)R^{-1}$, a straight-forward calculation yields

$$\begin{bmatrix} I & T \\ 0 & I \end{bmatrix} M(P) \begin{bmatrix} I & 0 \\ T^\top & I \end{bmatrix} = \begin{bmatrix} -\Lambda(P) & 0 \\ 0 & R \end{bmatrix},$$

where

$$\Lambda(P) = APA^\top - P + (\bar{C}^\top - APC^\top)R(P)^{-1}(\bar{C} - CPA), \tag{2.38}$$

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

$$\Lambda(P) \leq 0, \tag{2.39}$$

and

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

If P satisfies the algebraic Riccati equation

$$\Lambda(P) = 0, \tag{2.40}$$

$\operatorname{rank} M(P) = m$, and the corresponding spectral factor W is *square* $m \times m$. These P form a subfamily $\mathcal{P}_0$ in $\mathcal{P}$. For all square spectral factors, the condition (2.37) insures that all solutions of $DD^\top = R(P)$ must actually be square and invertible so that from the positive real equations (2.33) we obtain

$$B = (\bar{C}^\top - APC^\top)D^{-\top}, \tag{2.41}$$

If $P \notin \mathcal{P}_0$, W is rectangular and we can use instead a pseudo inverse of $D^\top$.

From spectral factors to stochastic realizations We now examine the implications of the spectral factorization results on state space realizations. Let W and $\bar{W}$ be two minimal stable and antistable spectral factors. It is intuitively clear that such factors play the role of transfer functions of "shaping filters" of the type (2.2) for the process $\mathbf{y}$. To make this precise however we need to manifacture two white noise processes $\mathbf{w}$ and $\bar{\mathbf{w}}$ serving as input white noise processes in the two filters, in such a way that the output process will be equal to $\mathbf{y}$. Naturally, the filter with unstable transfer function $\bar{W}$ represents in the time domain a convolution operator integrating the white input "backwards in time". When W and $\bar{W}$ are *square* and invertible transfer functions, the white noise processes can be generated by passing $\mathbf{y}$ through the inverse "whitening filters" W^{-1} and $\bar{W}^{-1}$. The idea here is just the same as the classical "whitening-shaping" filter dicotomy of Bode and Shannon [**?**]. The "whitened" processes $\mathbf{w}, \bar{\mathbf{w}}$ obtained in this way have in fact a flat spectral density and can be shown to be well-defined linear functionals of $\mathbf{y}$ (if the spectral factors have zeros on the unit circle this is however not trivial and requires the full power of spectral representation theory, see [67]).

In particular, since W_- is outer, the corresponding white noise process $\mathbf{w}_-$ is a *causal functional of* $\mathbf{y}$, i.e. $\mathbf{w}_-(t-1) \in H_t^-(\mathbf{y})$ for all t, so that we actually have $H_t^-(\mathbf{w}_-) = H_t^-(\mathbf{y})$. For this reason, $\mathbf{w}_-$ is called the (normalized) *forward innovation* process of $\mathbf{y}$ [73]. Similarly, the white noise process $\bar{\mathbf{w}}_+$ is an anticausal functional of $\mathbf{y}$, i.e. $\bar{\mathbf{w}}_+(t) \in H_t^+(\mathbf{y})$, so that $H_t^+(\bar{\mathbf{w}}_+) = H_t^+(\mathbf{y})$ for all t; $\bar{\mathbf{w}}_+$ is called the (normalized) *backward innovation* process of $\mathbf{y}$.

This picture generalizes also to all minimal (nonsquare) spectral factors. The only difficulty in the generalization is the nonuniqueness of the white generating noises $\mathbf{w}$ and $\bar{\mathbf{w}}$ associated to rectangular spectral factors. The difficulty can be

overcomed by selecting the input noises in a fixed ambient space, which is small enough to make the $\mathbf{w}$'s unique but also big enough to allow a solution $\mathbf{w}$ of the convolution equation $\mathbf{y} = W\mathbf{w}$ for each minimal spectral factor W (and $\bar{W}$). See [50].

Let us fix the minimal triplet $(A, C, \bar{C})$ in the representation (2.20) (or equivalently in (2.25)). We shall agree to parametrize all minimal stable spectral factors W by fixing the (A, C) parameters in a minimal realization (A, B, C, D). These are chosen the same (A, C) as in the factorization (2.20) for all W. The matrices (B, D) of each spectral factor are obtained by solving the positive real equations (2.33) in the given basis. Dually, we shall agree to represent any minimal antistable spectral factor $\bar{W}$ by the minimal realization $(A^\top, \bar{B}, \bar{C}, \bar{D})$, where $\bar{C}$ is given by (2.19) and $(\bar{B}, \bar{D})$ are obtained by solving the corresponding dual positive real equations.

Once the white noise inputs are defined by the appropriate whitening filters, it is obvious that the deterministic realizations of the spectral factors provide two families of minimal state-space stochastic realizations of the process $\mathbf{y}$, the first one being causal and the second anticausal. The state processes of the two realizations have as covariance matrices P and $\bar{P} = P^{-1}$, equal to the two solutions of the forward and dual LMI's. Equivalently, P is the unique solutions of the Lyapunov equation (2.7) and $\bar{P}$ of

$$\bar{P} = A^\top \bar{P} A + \bar{B}\bar{B}^\top$$

respectively. In this way we have *fixed the state processes* $\mathbf{x}$ *and* $\bar{\mathbf{x}}$ *of the two families of forward and backward minimal stochastic realizations of* $\mathbf{y}$. In fact the family of bases are fixed in such a way that each backward basis $\bar{\mathbf{x}}(t)$ at the instant of time t is the *dual basis* of some forward state $\mathbf{x}(t)$. We shall call this a **uniform choice of basis** in the family of all minimal Markovian splitting subspaces. It is not difficult to check that, a family of bases is uniform *if and only if each corresponding minimal stochastic realization of* $\mathbf{y}$ *leads to a representation of the output covariance by the* **same (minimal) triplet** $(A, C, \bar{C})$.

The steady-state Kalman filter realizations Let us fix the triplet $(A, C, \bar{C})$ of the realization (2.20) and let $\mathbf{w}_-$ and $\bar{\mathbf{w}}_+$ be the normalized white noise processes corresponding to the spectral factors W_- and W_+ as explained in the previous paragraph. The shaping filter W_- leads to a forward stochastic realization

$$\begin{cases} \mathbf{x}_-(t+1) = A\mathbf{x}_-(t) + B_-\mathbf{w}_-(t) \\ \mathbf{y}(t) = C\mathbf{x}_-(t) + D_-\mathbf{w}_-(t) \end{cases} \tag{2.42}$$

with state covariance P_-, and W_+ leads to a backward realization

$$\begin{cases} \bar{\mathbf{x}}_+(t-1) = A^\top \bar{\mathbf{x}}_+(t) + \bar{B}_+\bar{\mathbf{w}}_+(t-1) \\ \mathbf{y}(t-1) = \bar{C}\bar{\mathbf{x}}_+(t) + \bar{D}_+\bar{\mathbf{w}}_+(t-1) \end{cases} \tag{2.43}$$

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

These two stochastic realizations will play an important role in what follows. In fact, an important interpretation of these realizations is that,

Theorem 2.12. *The model* (2.42) *written in the form*

$$\mathbf{x}_-(t+1) = A\mathbf{x}_-(t) + B_-D_-^{-1}[\mathbf{y}(t) - C\mathbf{x}_-(t)]$$

is the steady-state Kalman filter of any minimal realization (2.1) of $\mathbf{y}$ *in the uniform choice of basis induced by the factorization* (2.20)*. In the same way*

$$\bar{\mathbf{x}}_+(t-1) = A^\top\bar{\mathbf{x}}_+(t) + \bar{B}_+\bar{D}_+^{-1}[\mathbf{y}(t-1) - C\bar{\mathbf{x}}_+(t)]$$

is the backward steady-state Kalman filter of all minimal backward realizations (2.16) in the uniform choice of basis induced by the factorization (2.20)*.*

A short geometric proof of this result is given in [**?**][Proposition 5.4.11]. Here we shall follow a more pedestrian approach which requires some basic facts about Kalman filtering which are recalled below. Consider the Kalman filter of a model (2.1)

$$\hat{\mathbf{x}}(t+1) = A\hat{\mathbf{x}}(t) + K(t)\mathbf{e}(t)$$

where $\hat{\mathbf{x}}(t) := \mathbb{E}\left[\mathbf{x}(t) \mid H_t^-(\mathbf{y})\right]$ is the state estimate and

$$\mathbf{e}(t) := \mathbf{y}(t) - C\hat{\mathbf{x}}(t)$$

is the innovation process. The Kalman gain is usually determined via the Riccati difference equation describing the evolution of the state error covariance matrix $Q(t) := \mathrm{Var}\{\mathbf{x}(t) - \hat{\mathbf{x}}(t)\}$,

$$Q(t+1) = AQ(t)A^\top - \left[AQ(t)C^\top + BD^\top\right]\left[CQ(t)C^\top + DD^\top\right]^{-1}\left[AQ(t)C^\top + BD^\top\right]^\top \tag{2.44}$$

with initial condition $Q(0) = P = \mathbb{E}\{\mathbf{x}(0)\mathbf{x}(0)^\top\}$. Because of orthogonality of the state estimation error to $\hat{\mathbf{x}}(t)$ the matrix $Q(t)$ is equal to $P - \hat{P}(t)$, where $\hat{P}(t)$ is the variance matrix of the state estimate. The Kalman gain is given by the formula

$$K(t) = \left[AQ(t)C^\top + BD^\top\right]\left[CQ(t)C^\top + DD^\top\right]^{-1}$$

Apparently all these expressions depend on P, B, D which vary with the particular realization (2.1). However subtracting (2.44) from the Lyapunov equation (2.7) determinig P, one can eliminate all these parameters and obtain an *invariant form* of the Riccati equation

$$\hat{P}(t+1) = A\hat{P}(t)A^\top - \left[\bar{C}^\top - A\hat{P}(t)C^\top\right]\left[\Lambda(0) - C\hat{P}(t)C^\top\right]^{-1}\left[\bar{C}^\top - A\hat{P}(t)C^\top\right]^\top \tag{2.45}$$

which has initial condition $\hat{P}(0) = 0$ and depends only on the parameters $(A, C, \bar{C}, \Lambda(0))$ of the output covariance. Similarly we get a model-independent expression for the Kalman gain

$$K(t) = \left[\bar{C}^\top - A\hat{P}(t)C^\top\right]\left[\Lambda(0) - C\hat{P}(t)C^\top\right]^{-1}.$$

Hence we have the following,

Proposition 2.13. *All minimal models* (2.1) *in the same uniform choice of basis have the same Kalman filter. Symmetrically, the backward Kalman filter estimating the state* $\bar{\mathbf{x}}(t)$ *of any backward model* (2.16)*, based on the future history* $H_t^+(\mathbf{y})$*, is the same if the backward models belong to the same uniform choice of basis.*

Having established this fact, we can now proceed with the proof of Theorem 2.12.

Proof. It is not hard to check that under our standing assumption on the models (2.1), the solution of the Riccati equation (2.44) converges as $t \to \infty$ to a positive semidefinite limit

$$Q(\infty) := P - \hat{P}(\infty) \geq 0.$$

In fact, in the limit $\hat{P}(\infty)$ must clearly satisfy the algebraic Riccati equation (2.40), and the inequality above just shows that $\hat{P}(\infty)$ is actually the minimal solution, P_-, of the LMI. In addition, in the steady state the Kalman gain can be expressed as

$$K(\infty) = \left[\bar{C}^\top - AP_-C^\top\right]\left[\Lambda(0) - CP_-C^\top\right]^{-1} = B_-D_-^\top(D_-D_-^\top)^{-1} = B_-D_-^{-1}$$

and $\mathrm{Var}\{\mathbf{e}_\infty(t)\} = D_-D_-^\top$. On the other hand it is a well-known fact that the "feedback matrix" $A - K(\infty)C$ appearing in the inverse of the steady state Kalman filter realization

$$\begin{array}{rcl} \hat{\mathbf{x}}(t+1) & = & [A - K(\infty)C]\hat{\mathbf{x}}(t) + K(\infty)\mathbf{y}(t) \\ \mathbf{e}_\infty(t) & = & -C\hat{\mathbf{x}}(t) + \mathbf{y}(t) \end{array}$$

(which actually is the steady-state whitening filter generating the innovation process $\mathbf{e}_\infty(t)$), has no eigenvalues outside of the unit circle (i.e. is causal) and therefore $\mathbf{e}_\infty(t)$ is obtained by filtering $\mathbf{y}$ by a causal filter. Hence the steady state Kalman filter realization is stable and its inverse is also stable. This means that its transfer function must be proportional to the outer spectral factor $W(z)_-$. In fact, the transfer function of the steady-state Kalman filter realization is $W_-(z)D_-^{-1}$ and the normalized steady-state innovation process $D_-^{-1}\mathbf{e}_\infty$ is the white noise $\mathbf{w}_-(t)$. □

Chapter 3
Realization by Canonical Correlation and Stochastic Balancing

We shall now concentrate on the construction of innovation realizations for a stochastic process $\mathbf{y}$. A particularly simple and useful method was introduced by Akaike [2] using the idea of *Canonical Correlation Analysis* (CCA).

3.1 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is an old concept in statistics [37]. Given two finite-dimensional subspaces $\mathbf{A}$, $\mathbf{B}$ of zero-mean random variables of dimensions n and m, one wants to find two special orthonormal bases say $\{\mathbf{u}_1, \dots, \mathbf{u}_n\}$ for $\mathbf{A}$, and $\{\mathbf{v}_1, \dots, \mathbf{v}_m\}$ for $\mathbf{B}$ such that

$$\mathbb{E}\{\mathbf{u}_k\, \mathbf{v}_h\} = \sigma_k\, \delta_{k,h}, \qquad k, h = 1, \dots, r \leq \min\{n, m\}$$

This is the same as asking that the correlation matrix of the two random vectors $\mathbf{u} := [\mathbf{u}_1, \dots, \mathbf{u}_n]'$ and $\mathbf{v} := [\mathbf{v}_1, \dots, \mathbf{v}_m]'$ made with the elements of the two bases, should be diagonal, i.e. assuming for example that $n \geq m$,

$$\mathbb{E}\{\mathbf{u}\mathbf{v}'\} = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots \\ \vdots & & \ddots & \\ & & & \sigma_m \\ 0 & 0 & \dots & 0 \;\; 0 \end{bmatrix}$$

That two orthonormal bases of this kind always exist follows by considering the following sequence of operations:

A**Algorithm 3.1.**

1. Pick any two basis vectors say $\mathbf{a}$ and $\mathbf{b}$ of the subspaces $\mathbf{A}$, $\mathbf{B}$. Let $\Sigma_{\mathbf{a}} := \mathbb{E}\{\mathbf{a}\mathbf{a}^\top\}$ $\Sigma_{\mathbf{b}} := \mathbb{E}\{\mathbf{b}\mathbf{b}^\top\}$ and $R := \mathbb{E}\{\mathbf{a}\mathbf{b}^\top\}$ be their variances and covariance matrices.

2. Orthonormalize $\mathbf{a}$ and $\mathbf{b}$ by say Gram-Schmidt. The orthonormal bases, $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ are obtained by factoring $\Sigma_{\mathbf{a}}$ and $\Sigma_{\mathbf{b}}$,

$$\Sigma_{\mathbf{a}} = L_{\mathbf{a}} L_{\mathbf{a}}^{\top}, \qquad \Sigma_{\mathbf{b}} = L_{\mathbf{b}} L_{\mathbf{b}}^{\top}$$

with $L_{\mathbf{a}}$ and $L_{\mathbf{b}}$ square and nonsingular and letting $\hat{\mathbf{a}} := L_{\mathbf{a}}^{-1}\mathbf{a}$, and $\hat{\mathbf{b}} := L_{\mathbf{b}}^{-1}\mathbf{b}$.

3. Let the covariance of the two orthonormal bases be $\hat{R} = \mathbb{E}\{\hat{\mathbf{a}}\hat{\mathbf{b}}^{\top}\} = L_{\mathbf{a}}^{-1} R L_{\mathbf{b}}^{-\top}$. Compute the SVD

$$\hat{R} = U \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^{\top}, \qquad \Sigma = \operatorname{diag}\{\sigma_1, \ldots, \sigma_r\}$$

where $\sigma_k > 0$ and $r = \operatorname{rank} R$.

4. Define

$$\mathbf{u} := U^{\top}\hat{\mathbf{a}}, \qquad \mathbf{v} := V^{\top}\hat{\mathbf{b}}$$

Then

$$\mathbb{E}\{\mathbf{u}\mathbf{v}^{\top}\} = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$$

So that the vectors $\mathbf{u}$ and $\mathbf{v}$ have the required properties.

To make this choice of bases unique one must requires that all the singular values σ_k' s, which are called **canonical correlation coefficients**, should be different (and ordered in decreasing magnitude). In this case the SVD in step (3) is in fact unique. The components $\{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ and $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ are called the **principal directions** (or the **canonical variables**) of $\mathbf{A}$ and $\mathbf{B}$. It follows from Schwartz inequality that all canonical correlation coefficients are bounded between zero and one, i.e.

$$1 \geq \sigma_k > 0$$

and have also an interpretation as *cosines of angles between subspaces*. We define: $\theta_k := \arccos \sigma_k$ to be **the $k-$th principal angle** of $\mathbf{A}$ and $\mathbf{B}$. It can be shown that $\sigma_{k+1} = \cos\theta_{k+1}$ is the solution of the following minimization problem

$$\sigma_{k+1} = \langle \mathbf{u}_{k+1}, \mathbf{v}_{k+1} \rangle = \max_{\mathbf{u}\in\mathbf{A},\, \mathbf{v}\in\mathbf{B}} \{\langle \mathbf{u}, \mathbf{v} \rangle\} \tag{3.1}$$

subject to:

$$\begin{aligned} &\langle \mathbf{u}, \mathbf{u}_h \rangle = 0 \quad h = 1, \ldots, k \\ &\langle \mathbf{v}, \mathbf{v}_h \rangle = 0 \quad h = 1, \ldots, k \\ &\|\mathbf{u}_h\| = \|v_h\| = 1 \end{aligned} \tag{3.2}$$

A proof can be found in [**?**, p.584].
Remark: Note that these concepts do not depend on the particular choice of bases $\mathbf{a}$, $\mathbf{b}$ initially made in the two subspaces.

Consider the orthogonal projection onto the subspace $\mathbf{A}$. The restriction of this operator to random variables of $\mathbf{B}$ is denoted $\mathbb{E}_{|\mathbf{B}}^{\mathbf{A}}$. One has

$$\mathbb{E}_{|\mathbf{B}}^{\mathbf{A}}\boldsymbol{\xi} = \sum_{k=1}^{n} \sigma_k \langle\, \boldsymbol{\xi}, \mathbf{v}_k \rangle\, \mathbf{u}_k = \begin{bmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_n \end{bmatrix} \Sigma \begin{bmatrix} \langle\, \mathbf{v}_1 \\ \vdots \\ \langle\, \mathbf{v}_n \end{bmatrix}, \boldsymbol{\xi} \rangle$$

so the canonical correlation coefficients can be viewed as *singular values of the restricted projection operator* $\mathbb{E}_{|\mathbf{B}}^{\mathbf{A}}$. This interpretation will be quite useful in the next section.

3.2 Canonical correlation and balanced stochastic realization

In this section we shall give a procedure for constructing the innovation models (2.42) using CCA. The key idea is to construct the state spaces X_- and X_+ of (**??**) and choose two special bases in them by doing CCA of the past and future spaces H^- and H^+ of $\mathbf{y}$. We shall essentially follow the same steps of the CCA algorithm presented in the previous section.

To this end it will first be useful to arrange past and future outputs as infinite vectors in the form,

$$\mathbf{y}_- = \begin{bmatrix} \mathbf{y}(-1) \\ \mathbf{y}(-2) \\ \mathbf{y}(-3) \\ \vdots \end{bmatrix} \qquad \mathbf{y}_+ = \begin{bmatrix} \mathbf{y}(0) \\ \mathbf{y}(1) \\ \mathbf{y}(2) \\ \vdots \end{bmatrix} \tag{3.3}$$

Let L_- and L_+ be the lower triangular Cholesky factors of the infinite block Toeplitz matrices

$$T_- := E\{\mathbf{y}_-\mathbf{y}_-^\top\} = L_-L_-^\top \qquad T_+ := E\{\mathbf{y}_+\mathbf{y}_+^\top\} = L_+L_+^\top$$

and let

$$\boldsymbol{\nu} := L_-^{-1}\mathbf{y}_- \qquad \bar{\boldsymbol{\nu}} := L_+^{-1}\mathbf{y}_+ \tag{3.4}$$

be the corresponding orthonormal bases in H^- and H^+ respectively. Define the *Hankel operator*:

$$\mathbb{H} := \mathbb{E}_{|H^+}^{H^-} \tag{3.5}$$

projecting the future (at time zero) of the process $\mathbf{y}$ orthogonally onto the past. Note that the image space of $\mathbb{H}$ is exactly the minimal Markovian splitting subspace X_-, the state space of the Kalman filter realization. Dually, consider the adjoint Hankel operator $\mathbb{H}^*$ which now projects the past H^- onto the future H^+. It is easy to see that

$$\mathbb{H}^* := \mathbb{E}_{|H^-}^{H^+} \tag{3.6}$$

so that the image space of the adjoint is X_+, the state space of the backward Kalman filter realization.

Proposition 3.1. *Let* $\mathbf{y}$ *be realized by a finite dimensional model of the form* (2.1)*. Then in the orthonormal basis* (3.4) *the matrix representation of the Hankel operator* $\mathbb{H}$ *is*

$$\hat{H}_\infty = L_+^{-1} H_\infty L_-^{-T} = L_+^{-1} \Omega \, \bar{\Omega}^\top L_-^{-T}, \tag{3.7}$$

where $H_\infty := \mathbb{E}\{\mathbf{y}_+ \mathbf{y}_-^\top\}$ *is the infinite Hankel matrix* (2.22) *and* Ω *and* $\bar{\Omega}$ *are the factors introduced in* (2.23)*, namely*

$$\Omega = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \end{bmatrix} \quad \textit{and} \quad \bar{\Omega} = \begin{bmatrix} \bar{C} \\ \bar{C}A^\top \\ \bar{C}(A^\top)^2 \\ \vdots \end{bmatrix}. \tag{3.8}$$

Proof. For any infinite string of real numbers b with finitely many nonzero components we shall write $\sum_k b_k \boldsymbol{\nu}_k := b^\top \bar{\boldsymbol{\nu}} \in H^+$. Recalling the formula (2.8), in view of (3.4),

$$E^{H^-} b^\top \bar{\boldsymbol{\nu}} = b^\top L_+^{-1} E\{\mathbf{y}_+ \mathbf{y}_-^\top\} (E\{\mathbf{y}_- \mathbf{y}_-^\top\})^{-1} L_- \boldsymbol{\nu}$$

and therefore it follows from $\mathbb{E}\{\mathbf{y}_- \mathbf{y}_-^\top\} = L_- L_-^\top$ that

$$\mathbb{H} \, b^\top \bar{\boldsymbol{\nu}} = b^\top \hat{H}_\infty \boldsymbol{\nu}$$

as claimed. □

Note that, with a uniform choice of bases, we obtain the same matrix factorization (2.23) for $\hat{H}_\infty$, irrespective of which X (i.e. which minimal realization of $\mathbf{y}$) is chosen.

Since $\mathbb{H}$ has finite rank n, it has exactly n nonzero singular values. The *canonical correlation coefficients* of the past and future paces H^- and H^+ of $\mathbf{y}$ can then be defined to be the (nonzero) singular values of $\mathbb{H}$. By Proposition 3.1 they are also the (nonzero) singular values of the "normalized" Hankel matrix H_∞ (which was the way they were introduced in the CCA algorithm of the previous section). Obviously we shall have $0 < \sigma_k < 1$. Some of the largest canonical correlation coefficients may be equal to one. This can however happen if and only if $H_- \cap H_+ \neq \{0\}$.

Observabilty and Constructibility Operators The observability ad constructibility operators associated to a Markovian splitting subspace X, are defined respectively as

$$\mathbb{O} : X \to H^+, \qquad \mathbb{O}\boldsymbol{\xi} := \mathbb{E}^{H^+} \boldsymbol{\xi} \tag{3.9}$$

$$\mathbb{C} : X \to H^-, \qquad \mathbb{C}\boldsymbol{\xi} := \mathbb{E}^{H^-} \boldsymbol{\xi} \tag{3.10}$$

The meaning is of best (minimum variance) estimators of the state given the future (or the past) of $\mathbf{y}$. They 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 $\mathbf{y}$, through the space X, as

$$\mathbb{H} = \mathbb{C}\mathbb{O}^* \tag{3.11}$$

a fundamental characterization of minimality being that X is a minimal splitting subspace[11] if and only if the factorization (3.11) is canonical, i.e. $\mathbb{C}$ is *injective* and $\mathbb{O}^* = \mathbb{E}^X|_{H^+}$ is a *surjective* operator. Hence, for a minimal state space X, both the *Gramians*[12] $\mathbb{C}^*\mathbb{C}$ and $\mathbb{O}^*\mathbb{O}$ are invertible maps $X \to X$.

Proposition 3.2. *Let* $\mathbf{x}$ *be a basis in a minimal state space* X *and* $\bar{\mathbf{x}}$ *be its dual basis. Then the matrix representations of the constructibility and observability Gramians relative to the bases* $\mathbf{x}$ *and* $\bar{\mathbf{x}}$ *respectively, are given by*

$$\mathbb{C}^{\hat{*}}\mathbb{C} = P_- \tag{3.12a}$$

$$\mathbb{O}^{\hat{*}}\mathbb{O} = \bar{P}_+ = P_+^{-1} \tag{3.12b}$$

where P_- *and* $\bar{P}_+$ *are the covariance matrices of* $\mathbf{x}_-$ *and* $\bar{\mathbf{x}}_+$ *in the given basis. Hence the Gramians do not depend on the particular minimal model chosen and are invariant over the family of all minimal realizations.*

Proof. Take any $\boldsymbol{\xi} \in X$ expressed in the basis $\mathbf{x}$ as $a^\top\mathbf{x}$. Since $\mathbb{C}$ projects onto the past, by definition of the Kalman filter state $\mathbf{x}_-$ we have,

$$\mathbb{C}(a^\top\mathbf{x}) = \mathbb{E}^{\mathbf{H}^-}(a^\top\mathbf{x}) = a^\top\mathbf{x}_-, \qquad a \in \mathbb{R}^n$$

and therefore

$$\langle \mathbb{C}(a^\top\mathbf{x}),\, \mathbb{C}(b^\top\mathbf{x})\rangle = \langle a^\top\mathbf{x},\, \mathbb{C}^*\mathbb{C}(b^\top\mathbf{x})\rangle = \langle a^\top\mathbf{x}_-, b^\top\mathbf{x}_-\rangle = a^\top P_- b.$$

On the other hand, $\langle a^\top\mathbf{x}, \mathbb{C}^*\mathbb{C}\, b^\top\mathbf{x}\rangle = \langle a,\, \mathbb{C}^{\hat{*}}\mathbb{C}\, b\rangle$, the last inner product being in $\mathbb{R}^n$. Hence (3.12a) follows. The other relation is proven by a dual argument. □

For a discussion of the meaning of the Gramians in a stochastic setting see [**?**]. Note that in a uniform choice of bases the models (2.1) and (2.16) with dual basis $\bar{\mathbf{x}} = P^{-1}\mathbf{x}$ yield the following representations

$$\mathbf{y}_- = \bar{\Omega}\bar{\mathbf{x}} + \text{terms in } H^-(\bar{\mathbf{w}}) \qquad \mathbf{y}_+ = \Omega\mathbf{x} + \text{terms in } H^+(\mathbf{w}) \tag{3.13}$$

from which it is easy to check that

$$\mathbb{E}\{\mathbf{y}_- \mid \bar{\mathbf{x}}\} = \bar{\Omega}\bar{\mathbf{x}}, \qquad \mathbb{E}\{\mathbf{y}_+ \mid \mathbf{x}\} = \Omega\mathbf{x}\,.$$

This means that the infinite matrices Ω and $\bar{\Omega}$ (operating from the left) are matrix representations of the adjoint operators $\mathbb{O}^*$ and $\mathbb{C}^*$ in the bases $\mathbf{x}$ and $\bar{\mathbf{x}}$ respectively. We shall call them the **extended observability and constructibility matrices**.

[11] i.e. is a minimal subspace making the future and past of $\mathbf{y}$ conditionally orthogonal given X.

[12] Note that the Gramians are finite dimensional operators, representable by $n \times n$ symmetric positive semidefinite matrices.

Incidentally, by substituting in (3.13) the formula for dual bases $\bar{\mathbf{x}} = P^{-1}\mathbf{x}$, we can formally compute also the orthogonal projections $\mathbf{x}_- = \mathbb{E}^{\mathbf{H}^-}\mathbf{x}$ and $\bar{\mathbf{x}}_+ = \mathbb{E}^{\mathbf{H}^+}\bar{\mathbf{x}}$ expressed in the bases $\mathbf{y}_-$ and $\mathbf{y}_+$, as

$$\mathbf{x}_- = \bar{\Omega}^\top T_-^{-1}\mathbf{y}_- \qquad \bar{\mathbf{x}}_+ = \Omega^\top T_+^{-1}\mathbf{y}_+. \tag{3.14}$$

Consequently, we have the the following explicit formulas for P_- and $\bar{P}_+$:

$$\bar{P}_+ = \Omega^\top T_+^{-1}\Omega \qquad P_- = \bar{\Omega}^\top T_-^{-1}\bar{\Omega}\,. \tag{3.15}$$

Now, recall that the squares of the nonzero singular values $\{\sigma_1, \sigma_2, \sigma_3, \ldots\}$ of the Hankel operator $\mathbb{H}$ are, by definition, the nonzero eigenvalues of the self adjoint operator $\mathbb{H}^*\mathbb{H}$. Namely

$$\mathbb{H}^*\mathbb{H}\boldsymbol{\xi}_k = \sigma_k^2\,\boldsymbol{\xi}_k, \qquad k = 1, \ldots, n$$

which in view of the factorization (3.11) can be written also as

$$\mathbb{O}^*\mathbb{O}\mathbb{C}^*\mathbb{C}(\mathbb{O}^*\boldsymbol{\xi}_k) = \sigma_k^2\,(\mathbb{O}^*\boldsymbol{\xi}_k), \qquad k = 1, \ldots, n$$

which involves the two Gramians. Combining this with Proposition 3.2 we get the following important result.

Theorem 3.3. *The squares of the (nonzero) canonical correlation coefficients of past and future of a stationary process admitting a minimal realization* (2.1) *are the eigenvalues of the product $P_-\bar{P}_+$. In formulas,*

$$\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\} = \lambda\{P_-\bar{P}_+\}. \tag{3.16}$$

This result should be compared with the expression of the singular values of (the Hankel matrix of) a *deterministic* system as the eigenvalues of the product of the reachabilty and observability Gramians. Inspired by the deterministic notion of a *balanced realization*, this suggests that an appropriate uniform choice of basis would be the one that makes P_- and $\bar{P}_+$ equal and equal to the diagonal matrix of nonzero canonical correlation coefficients.

Definition 3.4 (Desai and Pal). *A minimal realization $(A, C, \bar{C}, \Lambda(0))$ of a $m \times m$ positive real matrix is called* Stochastically Balanced[13] *if the minimal solutions P_-, $\bar{P}_+$ of the dual Linear Matrix Inequalities (2.32), (2.36) are both equal to the same diagonal matrix, i.e.*

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

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

Clearly, by Theorem 3.3 the σ_k' s must necessarily coincide with the canonical correlation coefficients of $\mathbf{y}$.

[13] Or *Positive-Real Balanced*.

Proposition 3.5. *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* $\{\sigma_1, \sigma_2, \ldots, \sigma_n\}$ *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).*

This can be accomplished by the following algorithm.

Algorithm for computing the change of basis matrix bringing a minimal positive-real realization $(A, C, \bar{C}, \Lambda(0))$ **to balanced 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

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

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

Balanced stochastic realization by CCA In view of Proposition 3.1, the infinite normalized Hankel matrix $\hat{H}_\infty$ is the matrix representation of the operator $\mathbb{H}$ in the orthonormal bases (3.4). Therefore in the singular-value decomposition

$$\hat{H}_\infty = U_\infty\Sigma_\infty V_\infty^\top = U\Sigma V^\top, \tag{3.17}$$

Σ is the diagonal $n \times n$ matrix consisting of the canonical correlation coefficients

$$\Sigma = \mathrm{diag}\{\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_n\}, \tag{3.18}$$

while Σ_∞ is the infinite matrix

$$\Sigma_\infty = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}.$$

Moreover U_∞ and V_∞ are infinite orthogonal matrices, and U and V are $\infty \times n$ submatrices of U_∞ and V_∞ with the the property that

$$U^\top U = I = V^\top V. \tag{3.19}$$

Mimiking the last step of Algorithm 3.1 we now rotate the the orthonormal bases (3.4) in H^+ and H^- to obtain $\mathbf{u} := U_\infty^\top\bar{\boldsymbol{\nu}}$ and $\mathbf{v} := V_\infty^\top\boldsymbol{\nu}$ respectively. Note

that $E\{\mathbf{u}\mathbf{v}^\top\} = \Sigma_\infty$. Therefore $\{\mathbf{v}_{n+1}, \mathbf{v}_{n+2}, \mathbf{v}_{n+3}, \ldots\}$ span the subspace of random variables in the past H^- which are orthogonal to the future, and likewise, $\{\mathbf{u}_{n+1}, \mathbf{u}_{n+2}, \mathbf{u}_{n+3}, \ldots\}$ span the subspace of random variables which are in the future but are orthogonal to the past H^-. In formulas

$$\overline{\text{span}}\{\mathbf{u}_{n+1}, \mathbf{u}_{n+2}, \mathbf{u}_{n+3}, \ldots\} = H^+ \cap (H^-)^\perp \tag{3.20}$$

$$\overline{\text{span}}\{\mathbf{v}_{n+1}, \mathbf{v}_{n+2}, \mathbf{v}_{n+3}, \ldots\} = H^- \cap (H^+)^\perp \,. \tag{3.21}$$

This leads to the following important characterization.

Theorem 3.6. *The canonical random variables* $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ *and* $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ *are orthonormal bases for the state spaces* X_+ *and* X_- *respectively, i.e.*

$$X_+ = \text{span}\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}, \qquad X_- = \text{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\} \tag{3.22}$$

Proof. This is true since $H^- = [H^- \cap (H^+)^\perp] \oplus X_-$ so that X_- is precisely the subspace of random variables in H^- having nonzero correlation with the future H^+ and, dually, since $H^+ = X_+ \oplus [H^+ \cap (H^-)^\perp]$ so that X_+ is the subspace of random variables in H^+ having nonzero correlation with the past H^-. Since $\{\mathbf{v}_{n+1}, \mathbf{v}_{n+2}, \mathbf{v}_{n+3}, \ldots\}$ and $\{\mathbf{u}_{n+1}, \mathbf{u}_{n+2}, \mathbf{u}_{n+3}, \ldots\}$ span $H^- \cap (H^+)^\perp$ and $H^+ \cap (H^-)^\perp$, respectively, the result follows. □

Now define the n-dimensional vectors

$$\bar{\mathbf{z}} = \begin{bmatrix} \sigma_1^{1/2}\mathbf{u}_1 \\ \sigma_2^{1/2}\mathbf{u}_2 \\ \vdots \\ \sigma_n^{1/2}\mathbf{u}_n \end{bmatrix} = \Sigma^{1/2} U^\top L_+^{-1} \mathbf{y}_+ \qquad \mathbf{z} = \begin{bmatrix} \sigma_1^{1/2}\mathbf{v}_1 \\ \sigma_2^{1/2}\mathbf{v}_2 \\ \vdots \\ \sigma_n^{1/2}\mathbf{v}_n \end{bmatrix} = \Sigma^{1/2} V^\top L_-^{-1} \mathbf{y}_- \tag{3.23}$$

From what we have seen before, $\mathbf{z}$ is a basis in X_- and $\bar{\mathbf{z}}$ is a basis in X_+, and they have the property that

$$E\{\mathbf{z}\mathbf{z}^\top\} = \Sigma = E\{\bar{\mathbf{z}}\bar{\mathbf{z}}^\top\}. \tag{3.24}$$

which means that they both belong to a stochastically balanced realization of $(A, C, \bar{C})$, as seen from the following statement.

Theorem 3.7 (L-P). *The basis vectors*

$$x_-(0) = \mathbf{z} \qquad x_+(0) = \bar{\mathbf{z}} \tag{3.25}$$

in X_- *and* X_+ *respectively belong to the same uniform choice of basis, i.e. to the same triplet* $(A, C, \bar{C})$*, and in this uniform choice*

$$P_- = \Sigma = \bar{P}_+. \tag{3.26}$$

If the canonical correlation coefficients $\{\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_n\}$ *are all distinct, this is, modulo similarity by a signature matrix* [14]*, the only choice in the equivalence class* $\{TAT^{-1}, CT^{-1}, T\bar{C}^\top;\ T \text{ nonsingular}\}$ *for which (3.26) holds.*

[14] A signature matrix is a diagonal matrix of ± 1.

Remark: Note, according to this theorem, that in the case of distinct canonical correlation coefficients, a stochastically balanced realization of $(A, C, \bar{C})$ defines a *canonical form* with respect to state space isomorphism by fixing the sign in, say, the first element in each row of C. Such canonical forms have been studied by Ober [**?**].

Proof. It follows from (3.7) and (3.17) that

$$E\{\bar{\mathbf{z}}\mathbf{z}^\top\} = \Sigma^2. \tag{3.27}$$

Now, choose $(A, C, \bar{C})$ so that $\bar{\mathbf{x}}_+(0) = \bar{\mathbf{z}}$, and let the bases in the other splitting subspaces be chosen accordingly so that the choice of bases is uniform. We want to show that $\mathbf{x}_-(0) = \mathbf{z}$. To this end, first note that $\mathbf{x}_+(0) = \Sigma^{-1}\bar{\mathbf{x}}_+(0)$ and that $\mathbf{x}_-(0) = E^{X_-}\mathbf{x}_+(0)$ by the Kalman filter property. Then, using formula (2.8) and the fact that $\mathbf{z}$ is a basis in X_-,

$$\mathbf{x}_-(0) = \Sigma^{-1}E\{\bar{\mathbf{z}}\mathbf{z}^\top\}\Sigma^{-1}\mathbf{z},$$

which, in view of (3.27), yields $\mathbf{x}_-(0) = \mathbf{z}$ as claimed. Hence (3.26) follows from (3.24).

Next, suppose that $(QAQ^{-1}, CQ^{-1}, \bar{C}Q^\top)$ is another uniform choice of bases which is also stochastically balanced. Since then $\mathbf{x}_-(0) = Q\mathbf{z}$ and, as is readily seen from the backward system (2.43) , $\bar{\mathbf{x}}_+(0) = Q^{-T}\bar{\mathbf{z}}$ so that $P_- = Q\Sigma Q^\top$ and $\bar{P}_+ = Q^{-T}\Sigma Q^{-1}$, (3.26) yields

$$Q\Sigma Q^\top = \Sigma \quad \text{and} \quad Q^{-T}\Sigma Q^{-1} = \Sigma,$$

from which we have

$$Q\Sigma^2 = \Sigma^2 Q.$$

Since Σ has distinct entries, it follows from [26, Corollary 2, p.223] that there is a scalar polynomial $\varphi(z)$ such that $Q = \varphi(\Sigma^2)$. Hence Q is diagonal and commutes with Σ so that, by $Q\Sigma Q^\top = \Sigma$, we have

$$QQ^\top = I.$$

Consequently, since Q is diagonal, it must be a signature matrix. □

Note that the normalization of H_∞ is necessary in order the singular values of H_∞ to become the canonical correlation coefficients, i.e., the singular values of $\mathbb{H}$. In fact, if we were to use the unnormalized matrix representation (2.22) of $\mathbb{H}$ instead, as may seem simpler and more natural, the transpose of (2.22) would not be the matrix representation of $\mathbb{H}^*$ in the same bases, a property which is crucial in the singular value decomposition above. This is because (2.22) corresponds to the bases $\mathbf{y}_-$ in H^- and $\mathbf{y}_+$ in H^+, which are not orthogonal. As we shall see in the next section, this holds also in applicable parts for the finite-dimensional case studied in Section 2, and therefore the normalized Hankel matrix $\hat{H}_\infty$, defined in Section **??**, is preferable to the unnormalized H_∞.

Formulas expressing the parameters $A, C, \bar{C}$ in terms of the chosen basis in the state space of a general minimal realization have been derived in (2.12) and (2.17). For any dual pair of bases $\mathbf{x}(0)$ and $\bar{\mathbf{x}}(0)$ we have

$$\begin{aligned} A &= \mathbb{E}\{\mathbf{x}(1)\mathbf{x}(0)^\top\}P^{-1} && (3.28a)\\ A^\top &= \mathbb{E}\{\bar{\mathbf{x}}(-1)\bar{\mathbf{x}}(0)^\top\}\bar{P}^{-1} && (3.28b)\\ C &= \mathbb{E}\{\mathbf{y}(0)\mathbf{x}(0)^\top\}P^{-1}, && (3.28c)\\ \bar{C} &= E\{\mathbf{y}(-1)\bar{\mathbf{x}}(0)^\top\}\bar{P}^{-1} = \mathbb{E}\{\mathbf{y}(-1)\mathbf{x}(0)^\top\} && (3.28d) \end{aligned}$$

Taking $\mathbf{x}_-(0) = \mathbf{z} = \Sigma^{1/2}V^\top\boldsymbol{\nu}$ we have $\mathbf{x}_-(1) = \Sigma^{1/2}V^\top\boldsymbol{\sigma}(\boldsymbol{\nu})$ where $\boldsymbol{\sigma}(\cdot)$ is the forward shift operator on random variables, i.e.

$$\boldsymbol{\sigma}(\boldsymbol{\nu}) = \boldsymbol{\sigma}\left(\begin{bmatrix}\boldsymbol{\nu}(-1)\\ \boldsymbol{\nu}(-2)\\ \boldsymbol{\nu}(-3)\\ \vdots\end{bmatrix}\right) := \begin{bmatrix}\boldsymbol{\nu}(0)\\ \boldsymbol{\nu}(-1)\\ \boldsymbol{\nu}(-2)\\ \vdots\end{bmatrix}.$$

Recalling that $P_- = \Sigma$, from (3.28) we get

$$A = \Sigma^{1/2}V^\top\mathbb{E}\{\boldsymbol{\sigma}(\boldsymbol{\nu})\boldsymbol{\nu}^\top\}V\Sigma^{1/2}\Sigma^{-1} \tag{3.29}$$

where, since $\boldsymbol{\nu}$ is an orthonormal process,

$$\mathbb{E}\{\boldsymbol{\sigma}(\boldsymbol{\nu})\boldsymbol{\nu}^\top\} = \begin{bmatrix}0 & 0 & 0 & \dots\\ I & 0 & 0 & \dots\\ 0 & I & 0 & \dots\\ 0 & 0 & I & \dots\\ \vdots & \vdots & \vdots & \dots\end{bmatrix} := \mathbf{S}$$

which is the matrix shifting one block downward. Hence $\mathbf{S}V = \begin{bmatrix}0\\ V\end{bmatrix}$ is the matrix having the first $m \times n$ block of zeros and the rest equal to the V matrix shifted downward by one block of m rows.

Theorem 3.8. *The stochastically balanced triplet $(\hat{A}, \hat{C}, \hat{\bar{C}})$ corresponding to the canonical bases $\mathbf{z}$ and $\bar{\mathbf{z}}$ can be expressed directly in terms of the SVD of the normalized Hankel matrix $\hat{H}_\infty$ of (3.7), as*

$$\begin{aligned} \hat{A} &= \Sigma^{1/2}V^\top\mathbf{S}\,V\Sigma^{-1/2}, && (3.30a)\\ \hat{A}^\top &= \Sigma^{1/2}U^\top\mathbf{S}\,U\Sigma^{-1/2}, && (3.30b)\\ \hat{C} &= \rho_1(H_\infty)L_-^{-\top}V\Sigma^{-1/2}, && (3.30c)\\ \hat{\bar{C}} &= \rho_1(H_\infty^\top)L_+^{-\top}U\Sigma^{-1/2}, && (3.30d) \end{aligned}$$

where $\rho_1(H_\infty)$ is the first block row of H_∞.

Proof. Formula (3.30) has already been proven. Dually, taking $\bar{\mathbf{x}}_+(0) = \bar{\mathbf{z}} = \Sigma^{1/2}U^\top\bar{\boldsymbol{\nu}}$, we have $\bar{\mathbf{x}}_+(-1) = \Sigma^{1/2}U^\top\boldsymbol{\sigma}^{-1}(\bar{\boldsymbol{\nu}})$ where $\boldsymbol{\sigma}^{-1}(\dot{})$ is now the backward shift operator on random variables so that the first block in $\boldsymbol{\sigma}^{-1}(\bar{\boldsymbol{\nu}})$ is $\bar{\boldsymbol{\nu}}(-1)$. From (3.28) recalling that in the current basis we also have $\bar{P}_+ = \Sigma$, we obtain,

$$\hat{A}^\top = \Sigma^{1/2}U^\top\mathbb{E}\{\boldsymbol{\sigma}^{-1}(\bar{\boldsymbol{\nu}})\bar{\boldsymbol{\nu}}^\top\}U\Sigma^{-1/2} = \Sigma^{1/2}U^\top\mathbf{S}U\Sigma^{-1/2},$$

which is (3.30).
Finally, taking again $\mathbf{x}(0) = \mathbf{z} = \Sigma^{1/2}V^\top\boldsymbol{\nu}$, (3.28) yields

$$\hat{C} = \mathbb{E}\{\mathbf{y}(0)\boldsymbol{\nu}^\top\}V\Sigma^{-1/2}\,.$$

Then, a symmetric argument, using (3.28), yields (3.30). □

Once $A, C, \bar{C}$ have been determined, to complete the conceptual stochastic realization procedure delineated in this section, we must explain how to compute the $(B_-,\, D_-)$ (or the $(B_+,\, D_+)$ parameters in the forward (or backward) innovation model (2.42) ((2.43)). This is however immediate at least in principle, as all we need to do is to compute the minimal symmetric solution P_- (resp. maximal symmetric solution P_+) of the Algebraic Riccati Equation

$$P = APA^\top + (\bar{C}^\top - APC^\top)R(P)^{-1}(\bar{C} - CPA^\top),$$

which both exist because of positivity of the covariance $\Lambda(t)$. From these, $B_\pm$ can be computed via (2.41).

Remark 3.1. We can make contact with the expressions for $(\hat{A}, \hat{C}, \hat{\bar{C}})$ which could be obtained by applying the Ho-Kalman realization algorithm starting from the factorization induced by the singular-value decomposition (3.17) of $\hat{H}_\infty$. Defining $\hat{\Omega}$ and $\hat{\bar{\Omega}}$ to be the extended observability and constructibility matrices in the canonical bases (3.25), namely

$$\hat{\Omega} = L_+U\Sigma^{1/2}, \qquad \hat{\bar{\Omega}} = L_-V\Sigma^{1/2} \tag{3.31}$$

so that

$$H_\infty = L_+U\Sigma^{1/2}\;\Sigma^{1/2}V^\top L_-^\top = \hat{\Omega}\,\hat{\bar{\Omega}}^\top\,,$$

we can recover the $\hat{A}$ matrix by imposing a recursive structure to $\hat{\Omega}$ according to which $\hat{A}$ is computed by solving the equation on the right side of

$$\hat{\Omega} = \begin{bmatrix} \hat{C} \\ (\downarrow \hat{\Omega}) \end{bmatrix}, \qquad (\downarrow \hat{\Omega}) = \hat{\Omega}\hat{A}$$

This i s called the *shift-invariance method* in the literature. Now since $(\downarrow \hat{\Omega}) = \mathbf{S}^\top\hat{\Omega}$, we can equivalently write

$$\mathbf{S}^\top\hat{\Omega} = \mathbf{S}^\top L_+U\Sigma^{1/2} = L_+U\Sigma^{1/2}\hat{A}$$

Now, since L_+ is the Cholesky factor of the block-symmetric infinite Toeplitz matrix T_+, we have the shift invariance property $\mathbf{S}^\top L_+ = L_+$ so that $\hat{A}$ can be found by solving

$$\mathbf{S}^\top U\Sigma^{1/2} = U\Sigma^{1/2}\hat{A} \tag{3.32}$$

which is exactly the same as (3.30). A dual argument, imposing a recursive structure to the constructibility matrix $\hat{\bar{\Omega}}$, yields (3.30). Similarly for $(\hat{C}, \hat{\bar{C}})$. In conclusion, *the expressions* (3.30) *for the system parameters are exactly the same one could obtain by applying the Ho-Kalman algorithm to the factorization* (3.17).

3.3 CCA realization based on Finite Data

In practice of course one never has an infinite string of covariances $\{\Lambda(t);\, t \geq 0\}$ but has available instead a finite string of observed data

$$\{y_0, y_1, y_2, \ldots, y_N\} \tag{3.33}$$

where, however, N may be quite large. More specifically, we assume that N is sufficiently large that replacing the ergodic limit for $N \to \infty$, of

$$\frac{1}{N+1}\sum_{t=0}^{N} y_{t+k}y_t^\top \qquad k \geq 0 \tag{3.34}$$

by truncated sums, yields good approximations of a suitable *finite* set of covariance lags, say

$$\{\Lambda(0), \Lambda(1), \ldots, \Lambda(T)\}, \tag{3.35}$$

where, of course, we need to bound T so that $T << N$ [15]. This is equivalent to saying that N is sufficiently large for

$$\frac{1}{N+1}\sum_{t=0}^{N} a^\top y_{t+k}y_{t+j}^\top b \tag{3.36}$$

to be essentially the same as the inner product $a^\top \mathbb{E}\{\mathbf{y}(k)\mathbf{y}(j)^\top\}b = a^\top\Lambda(k-j)b$ for arbitrary vectors $a, b \in \mathbb{R}^m$, provided $|k-j| \leq T$. Hence we may in our analysis proceed as if we had a finite sequence of random vectors

$$\{\mathbf{y}(0), \mathbf{y}(1), \mathbf{y}(2), \ldots, \mathbf{y}(T)\}, \tag{3.37}$$

extracted from the underlying stochastic process $\mathbf{y}$. For this reason in this section we shall proceed as if we had observations of $\mathbf{y}$ on the **finite interval** $[0,\, T]$.

Now, let us fix a "present" instant t, and partition the "data" (3.37) into two random vectors

$$\mathbf{y}_t^- = \begin{bmatrix} \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \\ \mathbf{y}(t-1) \end{bmatrix} \qquad \mathbf{y}_t^+ = \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{y}(t+1) \\ \vdots \\ \mathbf{y}(T-1) \end{bmatrix}, \tag{3.38}$$

[15] A practical rule of thumb is that one should take $T \simeq (1/50)N$ and should never take T greater than $(1/20)N$.

representing the past and the future at time t. We have saved the last value $\mathbf{y}(T)$ for later use, since in the following we shall need to consider also an *enlarged future string*

$$\bar{\mathbf{y}}_t^+ := \begin{bmatrix} \mathbf{y}_t^+ \\ \mathbf{y}(T) \end{bmatrix} .$$

We shall need to define the corresponding (finite-dimensional) subspaces Y_t^-, Y_t^+ and $\bar{Y}_t^+$ spanned by the scalar random variables components of $\mathbf{y}_t^-$, $\mathbf{y}_t^+$ and $\bar{\mathbf{y}}_t^+$ respectively.

In analogy to (3.76) introduce the (finite-interval) Hankel operator

$$\mathbb{H}_t := E^{Y_t^-}|_{Y_t^+}. \tag{3.39}$$

which has certainly a finite rank. It is easy to see that if $\mathbf{y}$ is representable by a finite dimensional realization (2.1) of dimension n, then rank $\mathbb{H}_t$ will in general be less or equal to n. The choice of the data size T and of the present time t are crucial for a correct determination of the rank and of the system order. A basic assumption in this respect will be the following.

Assumption 1. *The future and past horizons, $T-t$ and t, are chosen large enough so that*

$$\text{rank } \mathbb{H}_t = n \tag{3.40}$$

This condition is satisfied provided $T-t$ and t are (respectively) greater than the observability and constructibilty indices of the triplet $(A, C, \bar{C}^\top)$.

Next we shall consider (finite-interval) Markovian splitting subspaces for Y_t^- and Y_t^+, i.e., subspaces $\hat{X}_t$ for which $Y_t^- \vee \hat{X}_t^- \perp Y_t^+ \vee \hat{X}_t^+ \mid \hat{X}_t$, the past and future subspaces being relative to finite past and future temporal horizons. For any such subspace $\mathbb{H}_t$ admits a factorization

$$\mathbb{H}_t = \mathbb{C}_t \mathbb{O}_t^* \tag{3.41}$$

where the *constructibility and observability operators* relative to the finite past and future spaces Y_t^- and Y_t^+, are

$$\mathbb{O}_t : \hat{X}_t \to Y_t^+, \quad \mathbb{O}_t \boldsymbol{\xi} := \mathbb{E}^{Y_t^+} \boldsymbol{\xi}\,, \qquad \mathbb{C}_t : \hat{X}_t \to Y_t^-, \quad \mathbb{C}_t \boldsymbol{\xi} := \mathbb{E}^{Y_t^-} \boldsymbol{\xi}$$

the subspace $\hat{X}_t$ being *minimal splitting* if and only if these two operators are surjective (or equivalently if and only if the adjoints are injective) linear maps.

It is standard [49, **?**] to show that the *forward and backward predictor spaces*,

$$\hat{X}_-(t) := E^{Y_t^-} Y_t^+ \quad \text{and} \quad \hat{X}_+(t) = E^{Y_t^+} Y_t^-, \tag{3.42}$$

are such minimal Markovian splitting subspaces. The following proposition, whose proof can be found in [**?**], shows that any basis in $\hat{X}_-(t)$ is a Kalman filter estimate of the state of a minimal stationary model (2.1) and, dually, any basis in $\hat{X}_+(t)$

is an anticausal Kalman filter estimate of the state of a minimal stationary model (2.1)

Proposition 3.9 (L-P). *Let X be a (stationary) minimal Markovian splitting subspace for $\mathbf{y}$. Then $X_t := U^t X$ is also a minimal Markovian splitting subspace for Y_t^- and Y_t^+, and*

$$\hat{X}_-(t) = E^{Y_t^-} X_t \quad \textit{and} \quad \hat{X}_+(t) = E^{Y_t^+} X_t. \tag{3.43}$$

Conversely, let $\mathbf{y}$ admit minimal finite-dimensional realizations of the form (2.1). Then a basis $\hat{\mathbf{x}}(t)$ in $\hat{X}_-(t)$ has a representation

$$\hat{\mathbf{x}}(t) = E^{Y_t^-} \mathbf{x}(t) \tag{3.44}$$

where $\mathbf{x}(t) = U^t \mathbf{x}$ is a basis in X_t, the state space at time t of (2.1). For each fixed $\hat{\mathbf{x}}(t)$ and each choice of minimal stationary state space X, the basis $\mathbf{x}$ is unique. All $\mathbf{x}$ for which the representation (3.44) holds form a uniform choice of bases [16]. Dually, any basis $\hat{\bar{\mathbf{x}}}(t)$ in $\hat{X}_+(t)$ has a representation

$$\hat{\bar{\mathbf{x}}}(t) = E^{Y_t^+} \bar{\mathbf{x}}(t) \tag{3.45}$$

where $\bar{\mathbf{x}}(t) = U^t \bar{\mathbf{x}}$ is also a basis in X_t uniquely determined once X is fixed. All $\bar{\mathbf{x}}$ for which (3.45) holds also form a uniform choice of bases. If $\bar{\mathbf{x}}$ and $\mathbf{x}$ are dual bases then the minimal triplets $(A, C, \bar{C})$ and $(\bar{A}, \bar{C}, C)$ of the two uniform choice of bases, are the same.

Remark 3.2. According to the proposition above, for any minimal stationary state space X_t there is a unique pair $(\mathbf{x}(t), \bar{\mathbf{x}}(t))$ of bases such that $\hat{\mathbf{x}}(t)$ and $\hat{\bar{\mathbf{x}}}(t)$ admit the representations (3.44) and (3.45). If (and only if) $(\mathbf{x}(t), \bar{\mathbf{x}}(t))$ are dual bases then, by the last statement in the proposition, both $\hat{\mathbf{x}}(t)$ and $\hat{\bar{\mathbf{x}}}(t)$ are attached to the same triplet $(A, C, \bar{C})$. In this case we shall call them **uniform bases** [17]

It follows from this proposition that for any $\mathbf{x}$ in a uniform choice of bases, the random vector

$$\hat{\mathbf{x}}(t) = E^{Y_t^-} \mathbf{x}(t) \tag{3.46}$$

is invariant and its components form a basis in $\hat{X}_-(t)$. The vector $\hat{\mathbf{x}}(t)$ is the one-step predictor of $\mathbf{x}(t)$ based on Y_t^- and is then generated by the *transient Kalman filter*:

$$\hat{\mathbf{x}}(t+1) = A\hat{\mathbf{x}}(t) + K(t)[\mathbf{y}(t) - C\hat{\mathbf{x}}(t)]; \quad \hat{\mathbf{x}}(0) = 0, \tag{3.47}$$

where the gain $K(t)$ is given by

$$K(t) = (\bar{C}^\top - AP_-(t)C^\top)(\Lambda_0 - CP_-(t)C^\top)^{-1} \tag{3.48}$$

[16] and hence all corresponding stationary realizations determine the same minimal triplet $(A, C, \bar{C})$ parametrizing the covariance $\Lambda(\tau)$ of the process.

[17] If this is true, one can actually show that they are members of a uniform choice of bases in all *minimal finite-interval splitting subspaces*, as in the stationary setting.

and the filter estimate covariance

$$P_-(t) = E\{\hat{\mathbf{x}}(t)\hat{\mathbf{x}}(t)^\top\} \tag{3.49}$$

is the solution of the matrix Riccati equation

$$\begin{cases} P_-(t+1) = AP_-(t)A^\top + (\bar{C}^\top - AP_-(t)C^\top)(\Lambda_0 - CP_-(t)C^\top)^{-1}(\bar{C}^\top - AP_-(t)C^\top)^\top \\ P_-(0) = 0. \end{cases} \tag{3.50}$$

which are both invariant with respect to the choice of the model (2.1) (i.e. depend on the uniform choice of bases only)[18] .

Symmetrically, in terms of the backward system (**??**) corresponding to (2.1), the components of

$$\hat{\bar{\mathbf{x}}}(t) = E^{Y_t^+}\bar{\mathbf{x}}(t) \tag{3.51}$$

form a basis in $\hat{X}_+(t)$ and are generated by the backward Kalman filter

$$\hat{\bar{\mathbf{x}}}(t-1) = A^\top\hat{\bar{\mathbf{x}}}(t) + \bar{K}(t)[\mathbf{y}(t-1) - \bar{C}\hat{\bar{\mathbf{x}}}(t)]; \quad \hat{\bar{\mathbf{x}}}(T) = 0, \tag{3.52}$$

with

$$\bar{K}(t) = (C^\top - A^\top\bar{P}_+(t)\bar{C}^\top)(\Lambda_0 - \bar{C}P_-(t)\bar{C}^\top)^{-1} \tag{3.53}$$

where

$$\bar{P}_+(t) = E\{\hat{\bar{\mathbf{x}}}(t)\hat{\bar{\mathbf{x}}}(t)^\top\} \tag{3.54}$$

is obtained by solving the matrix Riccati equation

$$\begin{cases} \bar{P}_+(t-1) = A^\top\bar{P}_+(t)A + (C^\top - A^\top\bar{P}_+(t)\bar{C}^\top)(\Lambda_0 - \bar{C}\bar{P}_+(t)\bar{C}^\top)^{-1}(C^\top - A^\top\bar{P}_+(t)\bar{C}^\top \\ \bar{P}_+(T) = 0. \end{cases} \tag{3.55}$$

Now, it is well-known that both

$$\boldsymbol{\nu}(t) = (\Lambda_0 - CP_-(t)C^\top)^{-1/2}[\mathbf{y}(t) - C\hat{\mathbf{x}}(t)] \tag{3.56}$$

and

$$\bar{\boldsymbol{\nu}}(t) = (\Lambda_0 - \bar{C}\bar{P}_+(t)\bar{C}^\top)^{-1/2}[\mathbf{y}(t-1) - \bar{C}\hat{\bar{\mathbf{x}}}(t)] \tag{3.57}$$

are normalized white noises, called the forward respectively the backward *(transient) innovation processes*. Consequently, we may write the Kalman filter (3.47) as

$$\begin{cases} \hat{\mathbf{x}}(t+1) = A\hat{\mathbf{x}}(t) + B_-(t)\boldsymbol{\nu}(t) \\ \mathbf{y}(t) = C\hat{\mathbf{x}}(t) + D_-(t)\boldsymbol{\nu}(t) \end{cases} \tag{3.58}$$

where $D_-(t) := (\Lambda_0 - CP_-(t)C^\top)^{1/2}$ and $B_-(t) := K(t)D_-(t)$. Likewise, the backward Kalman filter (3.47) may be written

$$\begin{cases} \hat{\bar{\mathbf{x}}}(t-1) = A^\top\hat{\bar{\mathbf{x}}}(t) + \bar{B}_+(t)\bar{\boldsymbol{\nu}}(t-1) \\ \mathbf{y}(t-1) = \bar{C}\hat{\bar{\mathbf{x}}}(t) + \bar{D}_+(t)\bar{\boldsymbol{\nu}}(t-1) \end{cases} \tag{3.59}$$

[18]This invariance provides in fact a proof of the second half of the statement of Proposition 3.9.

where $\bar{D}_+(t) := (\Lambda_0 - \bar{C}\bar{P}_+(t)\bar{C}^\top)^{1/2}$ and $\bar{B}_+(t) := K(t)\bar{D}_+(t)$. Note that, since

$$P - P_-(t) = E\{[\mathbf{x}(t) - \hat{\mathbf{x}}(t)][\mathbf{x}(t) - \hat{\mathbf{x}}(t)]^\top\} \geq 0,$$

and, for the same reason, $\bar{P} - \bar{P}_+(t) \geq 0$, we have

$$P_-(t) \leq P \leq P_+(t) := \bar{P}_+(t)^{-1}, \tag{3.60}$$

so the prediction spaces $X_-(t)$ and $X_+(t)$ are extremal splitting subspaces, just as the stationary subspaces X_- and X_+ in (**??**).

Remark 3.3. *Comparing with* (2.1) *and* (2.16)*, we see that* (3.58) *and* (3.59) *are also stochastic realizations, which unlike* (2.1) *and* (2.16)*, are time-varying but where the* state process is now a function of the output variables $\{\mathbf{y}(t);\, t \in [0,\, T]\,\}$ only on the interval $[0, T]$. *In the stationary case the present state is instead a function of the* infinite past *(or of the* infinite future*) of the process. Hence the construction of these realizations only requires data which are actually available. Note also that in* (3.58)*,* (3.59) *the* $(A, C, \bar{C})$ *parameters* are the same of the stationary realization.

In complete analogy with the stationary framework in Section 3.2, we can express the $(A, C, \bar{C})$ parameters of the process by a finite-interval analog of formula (3.28), in terms of the state vectors of the transient realizations (3.58) and (3.59),

$$\begin{aligned}
A &= \mathbb{E}\,\{\hat{\mathbf{x}}(t+1)\hat{\mathbf{x}}(t)^\top\}P_-(t)^{-1} && \text{(3.61a)}\\
A^\top &= \mathbb{E}\,\{\hat{\bar{\mathbf{x}}}(t-1)\hat{\bar{\mathbf{x}}}(0)^\top\}\bar{P}_+(t)^{-1} && \text{(3.61b)}\\
C &= \mathbb{E}\,\{\mathbf{y}(t)\hat{\mathbf{x}}(t)^\top\}P_-(t)^{-1}, && \text{(3.61c)}\\
\bar{C} &= E\{\mathbf{y}(t-1)\hat{\bar{\mathbf{x}}}(t)^\top\}\bar{P}_+(t)^{-1} = \mathbb{E}\,\{\mathbf{y}(t-1)\hat{\mathbf{x}}(t)^\top\}\,. && \text{(3.61d)}
\end{aligned}$$

Hence, in order to compute the stationary $(A, C, \bar{C})$ matrices we need a procedure to construct appropriate basis vectors (i.e. states) in the predictor space $\hat{X}_-(t)$ and $\hat{X}_-(t+1)$. Similarly, we need a procedure to construct a basis (i.e. state) vector in the state spaces $\hat{X}_+(t)$ and $\hat{X}_+(t-1)$. This will be done next, by finite-interval CCA.

The canonical correlation coefficients

$$1 \geq \sigma_1(t) \geq \sigma_2(t) \geq \cdots \geq \sigma_n(t) > 0 \tag{3.62}$$

between the finite past Y_t^- and the finite future Y_t^+ are now defined as the singular values of the Hankel operator $\mathbb{H}_t$ defined by (3.41). To determine them we need a matrix representation of $\mathbb{H}_t$ in some orthonormal bases. Using the pair (3.56)–(3.57) of transient innovation processes for this purpose, we obtain a *normalized* matrix which, in analogy with (3.7), we shall denote $\hat{H}_t$, given by

$$\hat{H}_t = (L_t^+)^{-1} H_t (L_t^-)^{-\top} \tag{3.63}$$

where $H_t := \mathbb{E}\,\{\mathbf{y}_t^+\,(\mathbf{y}_t^-)^\top\}$ is a finite $m(T-t) \times mt$ upper left corner of the infinite Hankel matrix (2.22) and L_t^- and L_t^+ are the finite-interval counterparts of L_- and

L_+ respectively. The singular value decomposition yields

$$\hat{H}_t = U_t \Sigma_t V_t^\top, \tag{3.64}$$

where $U_t U_t^\top = I = V_t V_t^\top$, and Σ_t is the diagonal matrix of canonical correlation coefficients. Exactly as in Section 3.2 one can prove that

$$\begin{cases} \mathbf{z}(t) = \Sigma_t^{1/2} V_t^\top (L_t^-)^{-1} \mathbf{y}_t^- \\ \bar{\mathbf{z}}(t) = \Sigma_t^{1/2} U_t^\top (L_t^+)^{-1} \mathbf{y}_t^+ \end{cases} \tag{3.65}$$

are bases in $\hat{X}_-(t)$ and $\hat{X}_+(t)$ respectively and that

$$E\{\mathbf{z}(t)\mathbf{z}(t)^\top\} = \Sigma_t = E\{\bar{\mathbf{z}}(t)\bar{\mathbf{z}}(t)^\top\}. \tag{3.66}$$

In the present framework the invariance condition (3.16) becomes

$$\{\sigma_1(t)^2, \sigma_2(t)^2, \dots, \sigma_n(t)^2\} = \lambda\{P_-(t)\bar{P}_+(t)\}\,, \tag{3.67}$$

and, precisely as in the previous section, the canonical choice of bases (3.65) has the *finite-interval balancing property*, i.e.,

$$P_-(t) = \Sigma_t = \bar{P}_+(t), \qquad t \in [0,\, T] \tag{3.68}$$

Concerning the t-dependence it should be said that (3.65), as well as the relative variances $(P_-(t),\, \bar{P}_+(t))$, actually depend on the length of the past and future horizons $(t - t_0,\, T - t)$ (t_0 here is the initial time which was arbitrarily fixed equal to zero) and not on the specific "present date" t.

Now the question is to find bases in the *updated forward and backward predictor spaces*, which can be expressed as

$$\hat{X}_-(t+1) := E^{Y_{t+1}^-} \boldsymbol{\sigma} Y_t^+ \quad \text{and} \quad \hat{X}_+(t-1) = E^{Y_{t-1}^+} \boldsymbol{\sigma}^{-1} Y_t^-,$$

The CCA procedure followed in the stationary setting leads to consider, besides the Hankel matrix $H_t := \mathbb{E}\,\{\mathbf{y}_t^+(\mathbf{y}_t^-)^\top\}$ also two other finite block-Hankel matrices (note that they have different dimensions than H_t since $\mathbf{y}_{t+1}^-$ and $\bar{\mathbf{y}}_t^+$ are obtained by appending one more block row to $\mathbf{y}_t^-$ and $\mathbf{y}_t^+$)

$$H_{t+1} := \mathbb{E}\,\{\boldsymbol{\sigma}(\mathbf{y}_t^+)(\mathbf{y}_{t+1}^-)^\top\}, \qquad H_{t-1} := \mathbb{E}\,\{\mathbf{y}_t^-(\bar{\mathbf{y}}_t^+)^\top\} \tag{3.69}$$

The definition of H_{t-1} as $\mathbb{E}\,\{\boldsymbol{\sigma}^{-1}(\mathbf{y}_t^-)(\mathbf{y}_{t-1}^+)^\top\}$ may seem more natural but applying the backward shift to the vector $\mathbf{y}_t^-$ would require availability of $\mathbf{y}(-1)$ which we don't have. Introduce the SVD's of the corresponding normalized Hankel matrices

$$\hat{H}_{t+1} := U_{t+1}\Sigma_{t+1}V_{t+1}^\top, \qquad \hat{H}_{t-1} := U_{t-1}\Sigma_{t-1}V_{t-1}^\top\,. \tag{3.70}$$

Assuming for the moment that the diagonal matrices $\Sigma_t, \Sigma_{t+1}, \Sigma_{t-1}$ have the same number n of nonzero canonical correlation coefficients and hence, whitout loss of

generality could all be taken of dimension $n \times n$, one could generalize the updating formula of the stationary case by setting

$$\begin{cases} \mathbf{z}(t+1) = \Sigma_{t+1}^{1/2} V_{t+1}^{\top} (L_{t+1}^{-})^{-1} \mathbf{y}_{t+1}^{-} \\ \bar{\mathbf{z}}(t-1) = \Sigma_{t-1}^{1/2} U_{t-1}^{\top} (L_{t-1}^{+})^{-1} \mathbf{y}_{t-1}^{+} \,. \end{cases} \tag{3.71}$$

Unfortunately, while these formulas do provide bases in the updated predictor spaces they do not serve our purpose right. The explaination of this fact is in the following remark.

Remark 3.4. The main point of subspace identification is to recapture the *stationary* $(A, C, \bar{C})$ parameters of the process from the dynamic equations satisfied by the bases $\hat{\mathbf{x}}(t)$ and $\hat{\bar{\mathbf{x}}}(t)$ chosen in the finite-interval predictor spaces. As we have seen before these equations can be written as Kalman Filter recursions (3.58, 3.59) where $A, C, \bar{C}$ appear explicitly as dynamic parameters. However it should be stressed that the stationary parameters $A, C, \bar{C}$ appear in the Kalman-Filter equations simply because the bases $\hat{\mathbf{x}}(t)$ and $\hat{\bar{\mathbf{x}}}(t)$ have been shown to be obtainable by projection of the state $\mathbf{x}(t)$ of some stationary model of the process. In identification, where we are actually attempting to recover the stationary dynamics of $\mathbf{y}$, we do not have a stationary state-space model for $\mathbf{y}$ at our disposal. Indeed, it has taken us quite some work to understand how to pick bases in the predictor spaces $\hat{X}_{-}(t), \hat{X}_{-}(t+1)$ (and in the backward predictor counterparts) without using any information about the underlying stationary realizations. In particular, using the CCA method we have picked (3.65) and (3.71) just basing on a finite string of covariances of the process $\mathbf{y}$. It is however unclear at this point whether the difference equations relating these bases at time t and $t+1$ will actually involve the stationary parameters $A, C, \bar{C}$ we are after. In fact, the difference equations, although being of the Kalman Filter type seen so far, will not involve *constant* but rather *time varying* matrices $A, C, \bar{C}$ which will vary with t, the reason of this fact being that picking bases arbitrarily at different time instants cannot yield the same coefficient matrices (compare e.g. the formula $A = \mathbb{E}\{\mathbf{x}(t+1)\mathbf{x}(t)^{\top}\}\mathbb{E}\{\mathbf{x}(t)\mathbf{x}(t)^{\top}\}^{-1}$), even if we are dealing with the state equations of a stationary process. In fact, the time-varying $A, C, \bar{C}$ matrices wouldn't in general even be similar to the stationary parameters we are looking for. So the question arises of choosing bases $\hat{\mathbf{x}}(t)$ and $\hat{\bar{\mathbf{x}}}(t)$ at successive instants t and $t+1$ in such a way that their time evolution is described by difference equations with **constant** matrices $A, C, \bar{C}$. Bases $\hat{\mathbf{x}}(t)$ and $\hat{\mathbf{x}}(t+1)$ with this property will be called **coherent**. Dually, we have a concept of coherent bases $\hat{\bar{\mathbf{x}}}(t)$ and $\hat{\bar{\mathbf{x}}}(t-1)$. Only in this case the state space models yield $A, C, \bar{C}$ parameters which are a constant [19] minimal realization of the covariance of the stationary process $\mathbf{y}$.

The construction of coherent bases can be based on a recursive structure which links the three Hankel matrices H_t, H_{t+1}, H_{t-1}, which are all formed with the same

[19] "Constant" as opposed to possibly time varying realizations where, instead of (2.20), we could have $\Lambda(k) = C_k A_k^{k-1} \bar{C}_k^{\top}$ for $k > 0$.

(stationary) sequence $\{\Lambda(1), \Lambda(2), \ldots, \Lambda(T)\}$. To understand this structure we shall take a brief detour to review the partial realization problem.

3.4 Partial realization of covariance sequences

The **Minimal Partial Realization Problem** for the sequence (3.35) is of finding a minimal value of n and a minimal triplet of matrices $(A, C, \bar{C})$, of dimensions $n \times n$, $m \times n$ and $m \times n$ respectively, such that

$$CA^{i-1}\bar{C}' = \Lambda(i) \quad i = 1, 2, \ldots, T. \tag{3.72}$$

An infinite sequence

$$\{\Lambda_1, \Lambda_2, \Lambda_3, \ldots\} \tag{3.73}$$

is then obtained from (3.35) by setting $\Lambda_i := CA^{i-1}\bar{C}'$ for all i, in particular for all successive lags $i = T+1, T+2, \ldots$. Since $\Lambda_i := \Lambda(i)$ for $i = 1, \ldots, T$, this sequence is called a *minimal rational extension* of the finite sequence (3.35). The attribute "rational" is due to the fact that the elements of (3.73) are the coefficients of the Laurent expansion of the rational function

$$Z(z) = C(zI - A)^{-1}\bar{C}' = \Lambda_1 z^{-1} + \Lambda_2 z^{-2} + \ldots \tag{3.74}$$

about $z = \infty$. This is called a *rational extension of minimal degree* of (3.35).

Remark 3.5. *Even if we shall consistently refer to covariance matrices in view of future application to stochastic modeling, in this section we shall not be concerned with positivity questions and the partial realization problem could well be formulated in terms of an arbitrary finite sequence of matrices which are not necessarily covariances. Positivity is an issue which will be discussed separately later.*

In order to avoid trivial notational complications we shall assume that the index T is an even number say, $T = 2k$, and choose $t = k$ as the "middle point" of the interval $[0, T]$. Hence in this section we will be given certain $2k$ covariance matrices [20]

$$\{\Lambda(1), \ldots, \Lambda(2k)\}. \tag{3.75}$$

[20] These could be either the original data or the corresponding normalized covariances of the (finite-interval) whitened processes.

with which we shall form the block Hankel matrices

$$H_k := \begin{bmatrix} \Lambda(1) & \Lambda(2) & \cdots & \Lambda(k) \\ \Lambda(2) & \Lambda(3) & \cdots & \Lambda(k+1) \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda(k) & \Lambda(k+1) & \cdots & \Lambda(2k-1) \end{bmatrix} \tag{3.76}$$

$$H_{k+1} := \begin{bmatrix} \Lambda(1) & \Lambda(2) & \cdots & \Lambda(k) & \Lambda(k+1) \\ \Lambda(2) & \Lambda(3) & \cdots & \Lambda(k+1) & \Lambda(k+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Lambda(k) & \Lambda(k+1) & \cdots & \Lambda(2k-1) & \Lambda(2k) \end{bmatrix} = \begin{bmatrix} \Lambda(1) & \\ \Lambda(2) & \\ \vdots & \sigma H_k \\ \Lambda(k) & \end{bmatrix}$$

$$\bar{H}_{k+1} := \begin{bmatrix} \Lambda(1) & \Lambda(2) & \cdots & \Lambda(k) \\ \Lambda(2) & \Lambda(3) & \cdots & \Lambda(k+1) \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda(k) & \Lambda(k+1) & \cdots & \Lambda(2k-1) \\ \Lambda(k+1) & \Lambda(k+2) & \cdots & \Lambda(2k) \end{bmatrix} = \begin{bmatrix} \Lambda(1) & \Lambda(2) & \cdots & \Lambda(k) \\ & \sigma H_k & & \end{bmatrix} \tag{3.77}$$

where σH_k is the *shifted Hankel matrix*, of the same dimension of H_k but with all entries shifted by one time unit i.e. with $\Lambda(i+1)$ replacing $\Lambda(i)$ everywhere.

We quote from [68] the following uniqueness result of partial realizations.

Lemma 3.10. *The sequence (3.75) has a unique rational extension of minimal degree if and only if*

$$\text{rank}H_k = \text{rank}H_{k+1} = \text{rank}\bar{H}_{k+1} := n \tag{3.78}$$

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 (3.75), then there is a nonsingular* $n \times n$ *matrix* T *such that*

$$A_2 = T^{-1}A_1T, \quad C_2 = C_1T, \quad \bar{C}_2^\top = T^{-1}\bar{C}_1^\top. \tag{3.79}$$

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 reviewed below.

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

$$H_k = \Omega_k \bar{\Omega}_k^\top \tag{3.80}$$

where both factors Ω_k, $\bar{\Omega}_k$ have n linearly independent columns. Since by (3.78) $\text{columnspan}H_k = \text{columnspan}H_{k+1}$ and, dually, $\text{rowspan}H_k = \text{rowspan}\bar{H}_{k+1}$ there

exist matrices $\bar{C}, \bar{\Delta}, C, \Delta$ such that

$$\begin{bmatrix} \Lambda(1) \\ \Lambda(2) \\ \vdots \\ \Lambda(k) \end{bmatrix} = \Omega_k \bar{C}^\top, \quad \sigma H_k = \Omega_k \Delta \tag{3.81}$$

and

$$\begin{bmatrix} \Lambda(1) & \Lambda(2) & \cdots & \Lambda(k) \end{bmatrix} = C \bar{\Omega}_k^\top, \quad \sigma H_k = \Delta \bar{\Omega}_k^\top \tag{3.82}$$

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

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

In conclusion, the matrices

$$A = \Omega_k^{-L} \sigma H_k (\bar{\Omega}_k^\top)^{-R} \tag{3.83}$$

$$C = \begin{bmatrix} \Lambda(1) & \Lambda(2) & \cdots & \Lambda(k) \end{bmatrix} (\bar{\Omega}_k^\top)^{-R} \tag{3.84}$$

$$\bar{C} = \begin{bmatrix} \Lambda(1)^\top & \Lambda(2)^\top & \cdots & \Lambda(k)^\top \end{bmatrix} (\Omega_k^\top)^{-R} \tag{3.85}$$

are independent of the choice of the left- or right-inverses (denoted $^{-L}$ or $^{-R}$ respectively) and propagate the factorization (3.80) uniquely to H_{k+1} and $\bar{H}_{k+1}$ according to the formulas,

$$H_{k+1} = \begin{bmatrix} \Omega_k \bar{C}^\top & \Omega_k A \bar{\Omega}_k^\top \end{bmatrix} = \Omega_k \begin{bmatrix} \bar{C}^\top & A \bar{\Omega}_k^\top \end{bmatrix} := \Omega_k \, \bar{\Omega}_{k+1} \tag{3.86}$$

and

$$\bar{H}_{k+1} = \begin{bmatrix} C \bar{\Omega}_k \\ \Omega_k A \bar{\Omega}_k^\top \end{bmatrix} = \begin{bmatrix} C \\ \Omega_k A \end{bmatrix} \bar{\Omega}_k^\top := \Omega_{k+1} \, \bar{\Omega}_k^\top . \tag{3.87}$$

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

$$\Omega_{k+1} = \begin{bmatrix} C \\ \Omega_k A \end{bmatrix}, \quad \bar{\Omega}_{k+1} = \begin{bmatrix} \bar{C} \\ \bar{\Omega}_k A^\top \end{bmatrix} . \tag{3.88}$$

Now once (3.81, 3.82) 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 (3.88) can be interpreted as bona-fide recursions in k. From this we obtain precisely the classical structure of the observabililty and reconstructability matrices

$$\Omega_k = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{k-1} \end{bmatrix} \quad \bar{\Omega}_k = \begin{bmatrix} \bar{C} \\ \bar{C}A^\top \\ \vdots \\ \bar{C}(A^\top)^{k-1} \end{bmatrix}^\top , \tag{3.89}$$

seen in the literature.

It is important to note that under the equal ranks assumption (3.78), to each rank factorization (3.80) 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. We shall summarize the various steps in the following way.

Theorem 3.11. *Assume the rank condition* (3.78) *holds. Then each rank factorization (3.80) of the finite Hankel matrix H_k induces rank factorizations of H_{k+1} and $\bar{H}_{k+1}$ in* (3.86) *and* (3.87) *where the factors $\bar{\Omega}_{k+1}$ and Ω_{k+1} are uniquely determined. These factors satisfy the recursions* (3.88) *where the constant matrices $(A, C, \bar{C})$ are those uniquely determined by the factorization (3.80).*

The induced factorizations of H_{k+1} and $\bar{H}_{k+1}$ will be said to be **coherent** with that of H_k. Clearly coherent factorizations are unique.

Hankel factorization and choice of basis in the finite-interval predictor spaces

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

Proposition 3.12. *There is a one to one correspondence between rank factorizations* (3.80) *of the Hankel matrix H_k and choice of bases in the finite-interval predictor spaces $\hat{X}_-(k)$ and $\hat{X}_+(k)$. Given a rank factorization* (3.80) *the n-vectors*

$$\hat{\mathbf{x}}(k) := \bar{\Omega}_k^\top (T_k^-)^{-1} \mathbf{y}_k^-, \qquad \hat{\bar{\mathbf{x}}}(k) := \Omega_k^\top (T_k^+)^{-1} \mathbf{y}_k^+ \tag{3.90}$$

are uniform bases *in $\hat{X}_-(k)$ and $\hat{X}_+(k)$ in the sense defined in Remark 3.2. Conversely, given two such bases $\hat{\mathbf{x}}(k)$ and $\hat{\bar{\mathbf{x}}}(k)$, the matrices Ω_k and $\bar{\Omega}_k$ in the representations*

$$\mathbb{E}^{Y_k^-} \mathbf{y}_k^+ \,=\, \Omega_k \hat{\bar{\mathbf{x}}}(k), \qquad \mathbb{E}^{Y_k^+} \mathbf{y}_k^- \,=\, \bar{\Omega}_k \hat{\mathbf{x}}(k) \tag{3.91}$$

form a rank factorization of H_k of the type (3.80). *The factors Ω_k and $\bar{\Omega}_k$ are the extended observability and constructibility matrices of all minimal realizations of $\mathbf{y}$ in the uniform choice of bases determined by the triplet $(A, C, \bar{C})$, uniquely determined by the factorization* (3.80).

Proof. By (3.80),

$$\mathbb{E}^{Y_k^-} \mathbf{y}_k^+ = H_k (T_k^-)^{-1} \mathbf{y}_k^- = \Omega_k \bar{\Omega}_k^\top (T_k^-)^{-1} \mathbf{y}_k^- = \Omega_k \hat{\mathbf{x}}(k)$$

Since by definition the components of the projection of $\mathbf{y}_k^+$ onto Y_k^- span $\hat{X}_-(k)$, and since the columns of Ω_k are linearly independent it follows that $\hat{X}_-(k)$ is spanned by the scalar components of $\hat{\mathbf{x}}(k)$. These are also linearly independent since (as it is immediate to check) $\hat{\mathbf{x}}(k)$ has a positive definite variance matrix. A similar

reasoning for $\hat{\bar{\mathbf{x}}}(k)$ leads to the dual conclusion for $\hat{\bar{\mathbf{x}}}(\mathbf{k})$. Note that in (3.90) both bases are expressed in terms of the same parameters $(A, C, \bar{C})$ which are uniquely determined by the factorization (3.80). It follows that $\hat{\mathbf{x}}(k)$ and $\hat{\bar{\mathbf{x}}}(k)$ are uniform bases.

The converse, i.e. that the factorization of H_k follows from (3.91), is a consequence of the splitting property at time k of $\hat{X}_-(k)$ and $\hat{X}_+(k)$. In particular, let us look at

$$Y_k^+ \perp Y_k^- \mid \hat{X}_-(k)$$

which can be rewritten as,

$$\mathbb{E}\,\mathbf{y}(t)\mathbf{y}(s)^\top = \mathbb{E}\,\{\mathbb{E}\,[\mathbf{y}(t) \mid \hat{\mathbf{x}}(k)]\,\mathbb{E}\,[\mathbf{y}(s) \mid \hat{\mathbf{x}}(k)]^\top\}$$

for $t = k, \dots, 2k-1$ and $s = k-1, \dots, 0$. This relation arranged in matrix form is the same as $H_k = \Omega_k P(k) \tilde{\Delta}_k^\top$ where $P(k) := \mathbb{E}\,\hat{\mathbf{x}}(k)\hat{\mathbf{x}}(k)^\top > 0$ and $\tilde{\Delta}_k \hat{\mathbf{x}}(k) = \mathbb{E}\,[\,\mathbf{y}_k^- \mid \hat{\mathbf{x}}(k)]$. Letting $\bar{\Delta}_k := \tilde{\Delta}_k P(k)$ yields the factorization $H_k = \Omega_k \bar{\Delta}_k^\top$. 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 a dual full-rank factorization of H_k.

Let now $\hat{\mathbf{x}}_+(k) := \bar{P}(k)^{-1}\hat{\bar{\mathbf{x}}}(k)$ be the dual basis of $\hat{\bar{\mathbf{x}}}(k)$ and assume that $\hat{\mathbf{x}}(k)$ and $\hat{\bar{\mathbf{x}}}(k)$ are uniform bases, i.e. $\hat{\mathbf{x}}_+(k)$ is described by a (forward) model with the same parameters $(A, C, \bar{C})$ as $\hat{\mathbf{x}}(k)$. Then by Kalman filtering formulas, $\mathbb{E}\,[\hat{\mathbf{x}}_+(k) \mid \hat{\mathbf{x}}(k)] = \hat{\mathbf{x}}(k)$ (this is the time-varying analog of Proposition 2.13). Since the components of $\hat{\mathbf{x}}_+(k)$ belong to the future, we have $\mathbf{y}_k^- \perp \hat{\mathbf{x}}_+(k) \mid \hat{\mathbf{x}}(k)$, so that

$$\begin{aligned}\tilde{\Delta}_k P(k) &= \mathbb{E}\,\{\mathbb{E}\,[\,\mathbf{y}_k^- \mid \hat{\mathbf{x}}(k)]\,\hat{\mathbf{x}}(k)^\top\} = \mathbb{E}\,\{\mathbb{E}\,[\,\mathbf{y}_k^- \mid \hat{\mathbf{x}}(k)]\,\mathbb{E}\,[\,\hat{\mathbf{x}}_+(k) \mid \hat{\mathbf{x}}(k)]^\top\} \\ &= \mathbb{E}\,\{\,\mathbf{y}_k^-\hat{\mathbf{x}}_+(k)^\top\} = \mathbb{E}\,\{\,\mathbf{y}_k^-\hat{\bar{\mathbf{x}}}(k)^\top\}\bar{P}(k)^{-1} = \bar{\Omega}_k.\end{aligned}$$

which is what we needed to show. □

By an obvious extension of this result, bases in the predictor spaces $\hat{X}_-(t+1)$ and $\hat{X}_+(t-1)$ are biuniquely related to rank factorizations of the extended Hankel matrices H_{k+1} and H_{k-1} defined in (3.69). Note to this effect that

$$H_{k-1}^\top := \mathbb{E}\,\{\bar{\mathbf{y}}_k^+(\mathbf{y}_k^-)^\top\} = \bar{H}_{k+1} \tag{3.92}$$

Hence Proposition 3.12 has the following useful consequence.

Corollary 3.13. *Under the assumption* (3.78)*, coherent bases at time* $t+1$ *and* $t-1$ *correspond one-to-one to factorizations of the extended Hankel matrices* H_{k+1} *and* $\bar{H}_{k+1}$ *which are coherent with that of* H_k*.*

All of the above of course works, *mutatis mutandis*, for the normalized Hankel matrices $\hat{H}_k$, $\hat{H}_{k+1}$, $\hat{\bar{H}}_{k+1}$ and, in particular, for the *SVD-induced factorization* of $\hat{H}_k$ which is obtained by setting

$$\hat{\Omega}_k := U\Sigma_k^{1/2}, \qquad \hat{\bar{\Omega}}_k := V\Sigma_k^{1/2} \tag{3.93}$$

in (3.64). The moral of the story is that :

Fact : *The Hankel factorizations induced by the SVD's* (3.70) *are in general not coherent with the SVD-induced factorization* (3.64)*. Hence, the canonical bases in* $\hat{X}_-(t+1)$ *and* $\hat{X}_+(t-1)$ *given by* (3.65) *are* not coherent *with the canonical bases* (3.25) *at time* t *and do not yield a stationary realization of* $\Lambda(\tau)$*.*

Let

$$\hat{H}_{k+1} = \hat{\Omega}_k \, \hat{\bar{\Omega}}_{k+1}^\top, \qquad \hat{\bar{H}}_{k+1} = \hat{\Omega}_{k+1} \, \hat{\bar{\Omega}}_k^\top$$

be the factorizations of $\hat{H}_{k+1}$, $\hat{\bar{H}}_{k+1}$ coherent with (3.93), so that $\hat{\bar{\Omega}}_{k+1}^\top$ and $\hat{\Omega}_{k+1}^\top$ are uniquely determined by

$$\hat{\bar{\Omega}}_{k+1}^\top = \hat{\Omega}_k^{-L} \, \hat{H}_{k+1} = \Sigma_k^{-1/2} U^\top \hat{H}_{k+1}, \qquad \hat{\Omega}_{k+1}^\top = \hat{\bar{\Omega}}_k^{-L} \hat{H}_{k-1} = \Sigma_k^{-1/2} V^\top \hat{H}_{k-1} \, . \tag{3.94}$$

the last of which follows from the identity (3.92). The right way to choose bases in $\hat{X}_-(k+1)$ and $\hat{X}_+(k-1)$, coherent with (3.25), is to set,

$$\mathbf{z}(k+1) := \hat{\bar{\Omega}}_{k+1}^\top (L_{k+1}^-)^{-1} \mathbf{y}_{k+1}^- = \hat{\Omega}_k^{-L} \, \hat{H}_{k+1} (L_{k+1}^-)^{-1} \mathbf{y}_{k+1}^- \tag{3.95}$$

$$\bar{\mathbf{z}}(k+1) := \hat{\Omega}_{k+1}^\top (L_{k+1}^+)^{-1} \mathbf{y}_{k+1}^+ = \hat{\bar{\Omega}}_k^{-L} \hat{H}_{k-1} (L_{k+1}^+)^{-1} \mathbf{y}_{k+1}^+ \tag{3.96}$$

where we have used the expressions of $\hat{\bar{\Omega}}_{k+1}$ and $\hat{\Omega}_{k+1}$ given by (3.94). Hence *we actually don't need to compute SVD's of the updated Hankel matrices* $\hat{H}_{k+1}$, $\hat{\bar{H}}_{k+1}$.

Now recalling that the white noise vectors

$$\boldsymbol{\nu}_{t+1}^- := (L_{t+1}^-)^{-1} \mathbf{y}_{t+1}^-, \qquad \bar{\boldsymbol{\nu}}_{t-1}^+ := (L_{t-1}^+)^{-1} \mathbf{y}_{t-1}^+$$

have the same correlation structure of their stationary counterparts, namely

$$\mathbb{E}\{\boldsymbol{\nu}_{t+1}^- (\boldsymbol{\nu}_t^-)^\top\} = \mathbf{S}_t, \qquad \mathbb{E}\{\bar{\boldsymbol{\nu}}_{t-1}^+ (\bar{\boldsymbol{\nu}}_t^+)^\top\} = \bar{\mathbf{S}}_t$$

where now $\mathbf{S}_t$ and $\bar{\mathbf{S}}_t$ are finite shift matrices of dimensions $m(k+1) \times mk$ and $mk \times m(k+1)$, the statement of Theorem 3.8 can be modified to the finite-interval setting, in the following way.

Proposition 3.14. *The triplet* $(A, C, \bar{C})$ *corresponding to the (finite-interval) canonical bases* (3.65)*, is given by the formulas*

$$\hat{A} = \Sigma_k^{-1/2} U^\top \hat{H}_{k+1} \mathbf{S} V \Sigma_k^{-1/2}, \tag{3.97a}$$

$$\hat{A}^\top = \Sigma_k^{-1/2} V^\top \hat{H}_{k-1} \mathbf{S} U \Sigma_k^{-1/2}, \tag{3.97b}$$

$$\hat{C} = \rho_1(H_t)(L_k^-)^{-\top} V \Sigma_k^{-1/2}, \tag{3.97c}$$

$$\hat{\bar{C}} = \rho_1(H_t^\top)(L_k^+)^{-\top} U \Sigma_k^{-1/2}, \tag{3.97d}$$

where the operator $\rho_1(\cdot)$ *is extraction of the first* m*-dimensional row block as in Theorem 3.8.*

Because of uniqueness, these formulas must give the same triplet of matrices which one may compute by the (purely matrix-theoretic) *Shift Invariance Method*

which amounts to factoring the Hankel matrix H_k by SVD and solving the two sets of linear equations

$$\hat{\bar{\Omega}}_k^{-L}\hat{H}_{k-1} = \begin{bmatrix} \hat{C} \\ \hat{\Omega}_k\hat{A} \end{bmatrix}, \qquad \hat{\Omega}_k^{-L}\,\hat{H}_{k+1} \;=\; \begin{bmatrix} \hat{\bar{C}} \\ \hat{\bar{\Omega}}_k\hat{A}^\top \end{bmatrix} \tag{3.98}$$

Note that the left members in these equations are known quantities once the factors $\hat{\Omega}_k$, $\hat{\bar{\Omega}}_k$ have been computed. The triplet $(A, C, \bar{C})$ will be in finite-interval stochastically balanced form.

The final step of the finite-interval realization procedure is the computation of the steady state Kalman gain and stationary innovation variance or, equivalently, of the (B_-, D_-) parameters. Once $(A, C, \bar{C})$ are computed (here $\Lambda(0)$ is known) this is done exactly as in the stationary case by solving the ARE (2.40).

In the next chapter we shall apply this procedure directly to the observed time series.

Chapter 4
A Subspace Identification Algorithm for Time Series

The general idea of the so-called "Subspace methods" for identification of stochastic systems [60], is to substitute the partial realization approach with a geometric procedure operating directly on vector subspaces of $\mathbb{R}^N$ generated by the data. Our first duty will be to justify formally the equivalence of the two settings.

4.1 The Hilbert Space of a second-order ergodic time series

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

In this section we shall review the basic ideas behind this correspondence. For clarity of exposition we shall initially assume that $N = \infty$ and that the data

$$\{\dots, y_{-1}, y_0, y_1, \dots, y_t, \dots\} \tag{4.1}$$

have been collected starting from an infinitely remote time instant (so that the time series is actually doubly infinite).

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

$$\mathbf{y}(t) := [y_t, y_{t+1}, y_{t+2}, \dots] \tag{4.2}$$

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 $\mathbf{Y}$ of all finite linear combinations

$$\mathcal{Y} := \{\sum a_k^\top \mathbf{y}(t_k) \qquad a_k \in \mathbb{R}^m,\ t_k \in \mathbb{Z}\} \tag{4.3}$$

Note that the vector space $\mathcal{Y}$ is generated by the row spaces of the family of semi-

infinite matrices (4.2) or, equivalently is 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

$$\langle a^\top \mathbf{y}(k), b^\top \mathbf{y}(j) \rangle := \lim_{N\to\infty} \frac{1}{N+1} \sum_{t=0}^{N} a^\top y_{t+k} y_{t+j}^\top b = a^\top \Lambda_0(k-j) b, \tag{4.4}$$

Note that the limit exists because of our basic assumption of second-order ergodicity. In particular, Λ_0 is the true covariance function guaranteed by Wiener's Theorem in Section **??**. This inner product is then extended by linearity to all finite linear combinations of rows of the tail matrices (4.2), i.e. to the vector space $\mathbf{Y}$ which then becomes an inner product space. The inner product is nondegenerate if the Toeplitz matrix T_k, constructed with the true covariances $\{\Lambda_0(0), \Lambda_0(1), \ldots, \Lambda_0(k)\}$, 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 (4.4) $t = 0$ is replaced by an arbitrary initial instant t_0, so that

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

for all t_0 (wide-sense stationarity). We can also introduce a *shift operator* $\mathbf{U}$ on the family of semi-infinite matrices (4.2), by setting

$$\mathbf{U} a^\top \mathbf{y}(t) := a^\top \mathbf{y}(t+1) \quad t \in \mathbb{Z}, \quad a \in \mathbb{R}^m,$$

defining a linear map which is isometric with respect to the inner product (4.4) and extendable by linearity to all of $\mathbf{Y}$.

By closing the vector space $\mathcal{Y}$ with respect to convergence in the norm induced by the inner product (4.4), we obtain a Hilbert space of semi-infinite real sequences $\bar{\mathcal{Y}} = \text{closure}\{\mathcal{Y}\}$ to which the shift operator $\mathbf{U}$ can be extended by continuity as a unitary operator [21]

Now as stated formally in Proposition1.4 and in the subsequent generalization, we can always 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 $\Lambda_0(\cdot)$. Hence, at least as far as first and second order moments are concerned, the sequence of "tails" $\mathbf{y}$ defined in (4.2) 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

[21] Since we will not have much use for the completed space in the following, we shall not introduce a special symbol for it.

y provided we substitute expectations with ergodic limits of the type (4.4). Since we shall only be concerned with second order statistics in this paper, we may even formally *identify* the tail sequence **y** of (4.2) 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 $\mathbb{E}\{\xi\eta\}$ as being the (ergodic) inner product of the corresponding rows ξ and η. For reasons of uniformity of notation the inner product 4.4 will then be denoted

$$\langle \xi, \eta \rangle = \mathbb{E}\{\xi\eta\}, \tag{4.5}$$

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

Hence all defintions and results in the geometric stochastic Hilbert space framework of second order random variables introduced in Section **??** carry over unchanged to the present framework. The orthogonal projection of a (infinitely long) tail random variable $\boldsymbol{\xi}$ onto a subspace $\mathcal{X}$ of the space $\mathcal{Y}$ will still be denoted $\mathbb{E}[\,\xi \mid \mathcal{X}\,]$. Whenever $\mathcal{X}$ is given as the rowspace of some $n \times \infty$ matrix of generators $\mathbf{X}$, we shall write $\mathbb{E}[\,\boldsymbol{\xi} \mid \mathbf{X}\,]$ to denote the orthogonal projection expressed (perhaps nonuniquely) in terms of the generators. It is clear that for finitely generated subspaces we have the representation formula

$$\mathbb{E}[\,\boldsymbol{\xi} \mid \mathbf{X}\,] = \mathbb{E}(\boldsymbol{\xi}\,\mathbf{X}^\top)[\mathbb{E}(\mathbf{X}\,\mathbf{X}^\top)]^\dagger \mathbf{X} \tag{4.6}$$

where in case of linearly independent rows (i.e $\mathbf{X}$ of full row rank) we can substitute the pseudoinverse $\dagger$ with a true inverse. It is important to recognize that (4.6) is just the solution of the ordinary Least-Squares problem

$$\min_{a \in \mathbb{R}^n} \|\boldsymbol{\xi} - a^\top \mathbf{X}\|^2 = \min_{\mathbf{z} \in \mathcal{X}} \mathbb{E}[\,\boldsymbol{\xi} - \mathbf{z}\,]^2$$

A (stationary) stochastic realization of **y** is a representation of the type

$$\left\{ \begin{array}{lcl} \mathbf{x}(t+1) & = & A\mathbf{x}(t) + B\mathbf{w}(t) \\ \mathbf{y}(t) & = & C\mathbf{x}(t) + D\mathbf{w}(t) \end{array} \right. \tag{4.7}$$

where $\{\mathbf{w}(t)\}$ is p-dimensional normalized white noise , i.e. $\mathrm{E}\{\mathbf{w}(t)\mathbf{w}(s)^\top\} = I\delta_{ts}$ $\mathrm{E}\{\mathbf{w}(t)\} = 0$, etc..

Remark 4.1. It should be kept in mind that the various linear operations in (4.7) hold in the sense of the metric of the space $\bar{\mathcal{Y}}$ and are to be understood as "asymptotic equalities" between tail 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.

The geometric framework with finite data

For data of finite lenght N the inner product (4.5) must be approximated by a finite sum

$$\mathbb{E}\{\xi\eta\} \cong \frac{1}{N+1}\sum_{t=0}^{N}\xi_t\eta_t \tag{4.8}$$

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

Assume N is large enough for the time average in the ergodic limit (**??**) 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

$$\mathbf{Y}_k^- = \begin{bmatrix} \mathbf{y}(0) \\ \mathbf{y}(1) \\ \vdots \\ \mathbf{y}(k-1) \end{bmatrix} = \begin{bmatrix} y_0 & y_1 & \cdots & y_N \\ y_1 & y_2 & \cdots & y_{N+1} \\ \vdots & \vdots & & \vdots \\ y_{k-1} & y_k & \cdots & y_{k+N-1} \end{bmatrix} \tag{4.9}$$

$$\mathbf{Y}_k^+ = \begin{bmatrix} \mathbf{y}(k) \\ \mathbf{y}(k+1) \\ \vdots \\ \mathbf{y}(2k-1) \end{bmatrix} = \begin{bmatrix} y_k & y_{k+1} & \cdots & y_{k+N} \\ y_{k+1} & y_{k+2} & \cdots & y_{k+N+1} \\ \vdots & \vdots & & \vdots \\ y_{2k-1} & y_{2k} & \cdots & y_{2k+N-1} \end{bmatrix} \tag{4.10}$$

The relative rowspaces $\mathbf{Y}_k^-$, $\mathbf{Y}_k^+$ generated by the rows of the $m \times (N+1)$ matrices $\mathbf{y}(t)$ for $0 \le t < k$, and $k \le t < 2k$ 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[22].

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 $\mathbf{Y}_{[k,2k]}$. The present time k will be assumed large enough troughout. In particular we shall always assume that $k \ge n$ where n is the order of the underlying "true" system whenever it is needed.

The procedure propesed in [60] consists of a number of steps which can be described as follows.

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

[22] More accurate notations would be,

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

1. Form the sample finite-memory predictor matrices at time k, $\hat{\mathbf{Y}}_k^+ := \mathbb{E}^{\mathbf{Y}_k^-}\mathbf{Y}_k^+$ and at time $k+1$, $\hat{\mathbf{Y}}_{k+1}^+ := \mathbb{E}^{\mathbf{Y}_{k+1}^-}\mathbf{Y}_{k+1}^+$.

2. Compute the SVD

$$\hat{\mathbf{Y}}_k^+ = \begin{bmatrix}\hat{U}_k & \tilde{U}_k\end{bmatrix}\begin{bmatrix}\hat{\Sigma}_k & 0\\ 0 & \tilde{\Sigma}_k\end{bmatrix}\begin{bmatrix}\hat{V}_k^\top\\ \tilde{V}_k^\top\end{bmatrix} \tag{4.11}$$

and do order estimation by selecting the "significant" singular values, $\hat{\Sigma}_k$.

3. Extract the Observability matrix $\Omega_k := \hat{U}_k\hat{\Sigma}_k^{1/2}$ and the canonical basis in the (forward) predictor space: $\hat{\mathbf{X}}(k) := \hat{\Sigma}_k^{1/2}\hat{V}_k^\top$.

4. Pick the *coherent* basis in $\hat{X}_-(k+1)$

$$\hat{\mathbf{X}}(k+1) := \Omega_k^{-L}\hat{\mathbf{Y}}_{k+1}^+ = \hat{\Sigma}_k^{-1/2}\hat{U}_k^\top\hat{\mathbf{Y}}_{k+1}^+ \tag{4.12}$$

5. Compute (A, C) by solving

$$\begin{bmatrix}\hat{\mathbf{X}}(k+1)\\ \mathbf{Y}(k)\end{bmatrix} = \begin{bmatrix}A\\ C\end{bmatrix}\hat{\mathbf{X}}(k) + \begin{bmatrix}K(k)\\ I\end{bmatrix}\hat{\mathbf{E}}(k) \tag{4.13}$$

by Least Squares :

$$\min_{A,\,C} \| \begin{bmatrix}\hat{\mathbf{X}}(k+1)\\ \mathbf{Y}(k)\end{bmatrix} - \begin{bmatrix}A\\ C\end{bmatrix}\hat{\mathbf{X}}(k) \|$$

Or, Compute (A, C) by the shift-invariance method.

6. Compute $\bar{C}$ by

$$\bar{C} = \mathbb{E}\,\mathbf{Y}(k-1)\hat{\mathbf{X}}(k)^\top \tag{4.14}$$

7. Estimate Λ_0 by

$$\Lambda_0 = \mathbb{E}\,\mathbf{Y}(0)\mathbf{Y}(0)^\top \tag{4.15}$$

8. Attempt to solving the Algebraic Riccati equation $\Lambda(P) = 0$ and finding the unique stabilizing positive-definite solution P_-. If $(A, C, \bar{C}, \Lambda_0)$ is not positive real this attempt fails.

9. If $(A, C, \bar{C}, \Lambda_0)$ is not positive real re-run the algorithm with a larger k. Possibly the sample size may be too small; then increase N (if possible). If this does not work the data do not support a description by a linear model. Then stop.

10. If step (8) was successful, compute (B_-, D_-) by the formulas

$$D_- = (\Lambda_0 - CP_-C')^{1/2}, \quad B_- = (\bar{C}' - AP_-C')(\Lambda_0 - CP_-C')^{-1/2}. \tag{4.16}$$

The system-theoretical background of the procedure is exposed in the previous sections see in particular Theorem **??**.

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.

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 (3.80) since $\bar{\Omega}_{k+1}$ and Ω_k are uniquely determined from (3.80)

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

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

are equal to $\mathbb{E}\,[\mathbf{U}\mathbf{Y}_k^+|\bar{\mathbf{Y}}_{k+1}^-\,]$ 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

$$\hat{\mathbf{x}}(k+1) = (\Omega_k)^{-L}\tilde{\Omega}_k\mathbf{s}(k+1). \tag{4.17}$$

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.

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.

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 (**??**). For, choosing a basis $\hat{\mathbf{x}}(k)$ induces a rank factorization (3.80) where the matrix Ω_k is determined by $\hat{\mathbf{x}}(k)$ as shown in (3.91) of Theorem **??** above.

Alternative subspace Algorithm

1. Choose a basis $\hat{\mathbf{x}}(k)$ in $\hat{X}_{k-}$.
2. Compute the corresponding observability matrix Ω_k by (3.91).
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 (**??**).
5. Compute $(A, C, \bar{C})$ by the following formulas,

$$A = \mathbb{E}\,\hat{\mathbf{x}}(k+1)\hat{\mathbf{x}}(k)'\hat{P}(k)^{-1} \tag{4.18}$$

$$C = \mathbb{E}\,\mathbf{y}(k)\hat{\mathbf{x}}(k)'\hat{P}(k)^{-1} \tag{4.19}$$

$$\bar{C} = \mathbb{E}\,\mathbf{y}(k-1)\hat{\mathbf{x}}(k)' \tag{4.20}$$

where $\hat{P}(k) = \mathbb{E}\,\hat{\mathbf{x}}(k)\hat{\mathbf{x}}(k)' = \bar{\Omega_k}'(T_k^-)^{-1}\bar{\Omega_k}$

Note that (4.20) which formally is derived from the backward (or anticausal) 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,

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

Whether this reduced procedure can be turned into a more efficient numerical algorithm is however not clear as yet.

4.1.1 The least squares implementation

If the "expectation" operator $\mathbb{E}$ is written explicitely as in (4.8), then the formulas for $(A, C, \bar{C})$ of Theorem **??** express exactly the solution of the two dual *least squares problems,*

$$\min_{A,C} \| \begin{bmatrix} \hat{\mathbf{x}}(k+1) \\ \mathbf{y}(k) \end{bmatrix} - \begin{bmatrix} A \\ C \end{bmatrix} \hat{\mathbf{x}}(k) \|^2 \tag{4.21}$$

$$\min_{A',\bar{C}} \| \begin{bmatrix} \hat{\bar{\mathbf{x}}}(\mathbf{k-1}) \\ \mathbf{y}(k-1) \end{bmatrix} - \begin{bmatrix} A' \\ \bar{C} \end{bmatrix} \hat{\bar{\mathbf{x}}}(k) \|^2 \tag{4.22}$$

where the norm is 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 of Theorem **??** 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

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

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 know, is equivalent to assuming positivity of $(A, C, \bar{C})$ from the beginning.

4.1.2 Use of the SVD and the LQ factorization

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 (**??**) of the previous section leads to the choice

$$\Omega_k = L_k^+ U_k \Sigma_k^{1/2}, \quad \bar{\Omega}_k = L_k^- V_k \Sigma_k^{1/2} \tag{4.23}$$

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.
The LQ factorization a key step in subspace identification algorithms.

$$\begin{bmatrix} U \\ Y \end{bmatrix} = \begin{bmatrix} L_{uu} & 0 \\ L_{yu} & L_{yy} \end{bmatrix} \begin{bmatrix} Q_u^\top \\ Q_y^\top \end{bmatrix}$$

where $Q_u^\top Q_u = I,\ Q_y^\top Q_y = I,\ Q_u^\top Q_y = 0$ and L_{uu}, L_{yy} are lower triangular.

$$\begin{aligned} \mathbb{E}\left[Y \mid \mathcal{U}\right] &= Y Q_u \left[Q_u^\top Q_u\right]^{-1} Q_u^\top = L_{yu} Q_u^\top \\ \mathbb{E}\left[Y \mid \mathcal{U}^\perp\right] &= Y Q_y \left[Q_y^\top Q_y\right]^{-1} Q_y^\top = L_{yy} Q_y^\top \end{aligned}$$

$Q_y^\top$ an orthonormal basis for the orthogonal complement $\mathcal{U}^\perp$ in $\mathcal{U} \vee \mathcal{Y}$.

Bibliography

[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*, 1983, 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 realizatio 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 and S. Pinzoni, Acausal models and balanced realizations of stationary processes, *Linear Algebra and its Applications*, **205-206** (1994), 957-1003.

[67] N. I. Rozanov, *Stationary Random Processes*, Holden Day, 1963.

[68] A.Tether, Constructio of minimal state-variable models from input-output data, *IEEE Trans. Automatic Control* **AC-15** (1971), pp. 427-436.

[69] 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.

[70] 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.

[71] S. Weiland, Theory of Approximation and disturbance attenuation for linear systems *Doctoral Thesis*, University of Groningen, Jan 1991.

[72] N. Wiener, Generalized Harmonic Analysis, in *The Fourier Integral and Certain of its Applications*, Cambridge U.P. 1933.

[73] 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.

[74] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation,* IEEE Trans. Automatic Control **AC-16** (1971), pp. 621–634.

[75] 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.

[76] D.C. Youla, on The Factorization of Rational Matrices, *IRE Transactions PGIT* , **7**: 1961.