High-level primitives for recursive maximum likelihood estimation

Bernard C. Levy, Albert Benveniste, Ramine Nikoukhah

High-level Primitives for Recursive Maximum Likelihood Estimation

Bernard C. Levy

Albert Benveniste

Ramine Nikoukhah

N° 2088

*R**apport
de recherche*

1994

High-level Primitives for Recursive Maximum Likelihood Estimation[*]

Bernard C. Levy[**]

Albert Benveniste[***]

Ramine Nikoukhah[****]

Abstract: This paper proposes a high level language constituted of only a few primitives and macros for describing recursive maximum likelihood (ML) estimation algorithms. This language is applicable to estimation problems involving linear Gaussian models, or processes taking values in a finite set. The use of high level primitive allows the development of highly modular ML estimation algorithms based on only few numerical blocks. The primitives, which correspond to the combination of different measurements, the extraction of sufficient statistics, and the conversion of the status of a variable from unknown to observed, or vice-versa, are first defined for linear Gaussian relations specifying mixed deterministic/stochastic information about the system variables. These primitives are used to define other macros, and are illustrated by considering the filtering and smoothing problems for linear descriptor systems. In a second stage, the primitives are extended to finite state processes, and are used to implement the Viterbi ML state sequence estimator for a hidden Markov model.

Key-words: Maximum Likelihood Estimation, Descriptor Systems, Failure Detection, Hidden Markov Models, Viterbi Algorithm.

(Résumé : tsvp)

Primitives de haut niveau pour l'estimation récursive

Résumé : On présente un langage de haut niveau constitué d'un petit nombre de primitives et de macros pour décrire des algorithmes récursifs d'estimation au maximum de vraisemblance. Ce langage est applicable au cas linéaire Gaussien, mais aussi au cas de variables à valeurs dans un alphabet fini.

Mots-clé : Estimation au maximum de vraisemblance, systèmes descripteurs, détection de pannes, modèles de Markov cachés, Viterbi.

1 Introduction

In spite of the fact that Kalman filtering relies on a simple Gram-Schmidt orthogonalization principle, over the years, the literature devoted to Kalman filtering and smoothing [1], [2] and its square-root or fast algorithms implementations [3], [4] has become relatively complex. To deal with numerical conditioning problems, such the possible singularity of the measurement noise, or large uncertainties in the initial state variance, a number of variants of the basic filtering and smoothing algorithms have been developed. Although most of these algorithms are superficially similar, they usually require distinct implementations, which results in duplication of programming effort and a general state of confusion for the end user. There is therefore a need to develop a modular environment, where all algorithms can be constructed from a few basic modules, and where the underlying simplicity of all filtering and smoothing procedures becomes again transparent.

One could of course argue that numerical signal processing software packages, such as MATLAB, provide the desired environment for developing estimation algorithms in a modular manner. However, the functions performed by MATLAB are of a low level nature, and do not capture the statistical aspects of operations arising in recursive maximum likelihood (ML) estimation algorithms. By comparison, certain real-time signal processing languages, such as SIGNAL [5] or its extension to random processes called SIGNalea [6], employ only a small number of high-level primitives, which admit simple interpretations in terms of operations on asynchronous data streams. While our goal here is more circumscribed than for SIGNAL, since we focus exclusively on estimation problems, and do not have to contend with timing issues, we shall retain the idea of employing just a few primitives to describe recursive estimation algorithms in a modular format. The objects on which the primitives operate are called *observations*, where an "observation" describes the hard constraints and probabilistic relationship existing between a set of measurements and unknown variables. The primitives and macros that we construct correspond in fact to simple statistical operations, such as the combination of observations, the extraction of sufficient statistics of minimal dimension, or the computation of marginal or conditional densities.

The first step towards the development of estimation primitives was taken in [7], where two operations, called reduction and extraction, were introduced to formulate linear estimation problems over graphs. The reduction operation compresses redundant observations, whereas an extraction has the effect of removing from an observation some variables which are no longer of interest. Unfortunately, the construction of the reduction and extraction operations given in [7] is nonprobabilistic and focuses only on the generation of ML estimates and error variances. As a consequence, these two operations alone cannot accomplish certain tasks, such as the generation of innovations or the computation of the likelihood function of an observation sequence. In this paper, we adopt a probabilistic point of view towards the construction of primitives. Since the statistical concepts employed to define the primitives are quite general, we do not restrict our attention to Gaussian processes. We consider all ML estimation problems that are amenable to finite computer programs, i.e. those for which conditional densities can be finitely parametrized. In addition to Gaussian models,

these include processes taking values in a finite set, such as finite-state Markov chains [8] or Markov random fields [9].

As in [7], an important feature of the approach we employ to formulate recursive ML estimation problems is that we make no distinction between equations describing the model dynamics and measurements. In the case of dynamics, observations may involve several successive states of the process. This viewpoint is motivated by the observation that among all the equations describing a certain stochastic process, the division between a subset employed to construct the process, and another measurement subset is totally artificial. What really matters is whether, given all the available dynamics and measurements, we can estimate the variables which are of interest. This raises two issues. The first concerns the fact that any dynamic or measurement equation contains in general a mixture of deterministic and probabilistic information about the system variables. More precisely, depending on whether noise enters in an equation or not, this equation specifies either a deterministic algebraic constraint or a probability density or distribution for the system variables. This suggests that the proper probabilistic framework for modeling systems is to consider probability densities or distributions defined over submanifolds (linear subspaces in the Gaussian case) of the underlying domain. This feature appears already in SIGNalea[6], where stochastic processes are constructed by combining deterministic relations with a random variable generator. The fact that we process all dynamics and measurements progressively, instead of constructing the process a priori, engenders a second problem. Specifically, it is possible that certain variables which are estimable from the complete observation set, may remain unknown on the basis of the current partial observation set. Thus, we are confronted with a situation where at any given time we have mixed information about the system variables: some are known exactly, others are unknown, and the remainder admits a probability distribution.

The paper is organized as follows. We consider first the case of linear Gaussian models. In Section 2 we show that we can associate to a linear Gaussian observation a pair formed by a constraint space for the system parameters and measurements, and a Gaussian probability density defined over this space. A criterion is presented to determine whether a variable is estimable from the specification of such a pair, and the maximum likelihood and least-squares estimates are obtained by solving constrained optimization problems. The high-level ML estimation primitives are introduced in Section 3. The first primitive performs the combination of two observations. This involves imposing to the combined observation the deterministic constraints associated to each of its components. Another difficulty arises when the observations we combine contain common parameters or measurement vectors. The equality of these parameters and measurements must be taken into account in the combination step. The second and third primitive are formed by the reduction operation and its complement, the coreduction. From a statistical point of view, the reduction of an observation generates a minimal sufficient statistic while preserving the constraint space of the parameter vector. The coreduction extracts from an observation the information concerning only its measurement vector. It yields the likelihood function of the measurement vector, as well as the deterministic constraints, often called parity checks, that must be applied to the measurement to determine whether it is compatible with the observation

model. The last primitive, called mutation, is needed to define conditional observations, and consists of switching the status of a parameter or measurement from unknown to known, or vice-versa. In our framework, the extraction operation of [7] is only a macro, which can be implemented easily in terms of the four primitives. Efficient numerical implementations of the primitives are also provided, so that to each primitive corresponds both a statistical concept and a numerical module.

The primitives and macros are used to obtain a general recursion principle for the computation of ML estimates, and the evaluation of the likelihood function and parity checks of a stream of observations. This principle is illustrated in Section 4 by considering the filtering and smoothing of linear Gaussian descriptor systems, as well as a simple failure detection problem. High level programs are provided for the Kalman filter, and the two-filter and Rauch-Tung-Striebel smoothers. These programs are so compact it is hard to believe they implement a complete Kalman filter or smoother. However, by keeping track of the step by step implementation of each primitive, it is easy to relate them to standard Kalman filter or smoother implementations.

The case of a parameter process taking values in a finite set is considered in Section 5. Depending on whether the measurements are continuous or discrete, an observation is modeled as a probabilility density or distribution defined over a constraint manifold, corresponding to the deterministic constraints imposed on the parameters and measurement. Because of the discrete nature of the parameter process, only ML estimates are of interest. Also, since the primitives correspond to general statistical operations, it is not difficult to adapt them to the given model. The only difference is that depending on whether the primitives act on Gaussian or finite-valued variables, they rely on different numerical algorithms. The applicability of the primitives is illustrated by considering the maximum likelihood state sequence estimation problem for a hidden Markov model. By realizing that this problem has the same graph dependency structure between observations and parameters as the linear descriptor estimation problem of Section 4 (see [7] and Section 5 for a discussion of the XO-graph representation of an estimation problem), we demonstrate, somewhat surprisingly, that the Viterbi algorithm admits the same high-level program as the double-sweep Rauch-Tung-Striebel smoother. Thus, although the two algorithms look very different, the use of primitives sensitive to data type allows us to show they are identical, which makes precise an analogy observed in [10].

2 Estimation over Constrained Spaces

A Observation Model

Consider a linear observation

$$o \quad : \quad Ey = Ax + Bu \,, \tag{2.1}$$

where $u \sim \mathcal{N}(0, I_m)$ is a zero-mean Gaussian vector of $\mathbf{R}^m$, and the measurement $y \in \mathbf{R}^p$ and parameter vector $x \in \mathbf{R}^n$ are respectively known (we shall say *observed*) and *unknown*.

The observation o has dimension q, where q denotes the dimension of the vector Ey, so that the matrices E, A and B have sizes $q \times p$, $q \times n$ and $q \times m$, respectively. In (2.1), the covariance BB^T of the noise Bu need not be invertible, but without loss of generality, it can be assumed that B is one to one, i.e., B has full column rank.

The observation model (2.1) is nonstandard, since it is customary to set $E = I$, with $q = p$. However, the greater generality afforded by the model (2.1) will enable us to combine observations o_1 and o_2 whose measurement vectors y_1 and y_2 contain common components. This model allows also the existence of algebraic constraints between the entries of y. In this context, because no constraint is imposed on the rows of E, it is worth noting that the distribution $u \sim \mathcal{N}(0, I_m)$ is an *a-priori* distribution for the noise, before the specification of equation (2.1). After the observation (2.1) becomes known, some components of u may be known exactly. The following lemma describes more precisely the information provided by equation (2.1).

Lemma 2.1 (canonical decomposition)

1) *Given a linear observation o of the form (2.1), there exists a transformation (S, V) with S invertible and V orthonormal such that*

$$
SA = \begin{bmatrix} M_1 \\ M_2 \\ 0 \\ 0 \end{bmatrix} \qquad SBV^T = \begin{bmatrix} I & 0 \\ 0 & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix} \tag{2.2}
$$

with $M = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}$ onto, so that if

$$
SE = F = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} \qquad Vu = \begin{bmatrix} u_1 \\ u_3 \end{bmatrix}, \tag{2.3}
$$

the observation o takes the form

$$
\begin{bmatrix} F_1 \\ F_2 \\ F_3 \\ F_4 \end{bmatrix} y = \begin{bmatrix} M_1 \\ M_2 \\ 0 \\ 0 \end{bmatrix} x + \begin{bmatrix} u_1 \\ 0 \\ u_3 \\ 0 \end{bmatrix}, \tag{2.4}
$$

where $u_1 \sim \mathcal{N}(0, I)$ and $u_3 \sim \mathcal{N}(0, I)$ are independent.

2) *The canonical decomposition (2.2)–(2.4) provides the following information about x and y. First, $[y^T \ x^T]^T$ belongs to the subspace $\mathcal{C}$ of $\mathbf{R}^{p+n}$ defined by the two constraints*

$$
\begin{aligned}
F_2 y &= M_2 x \tag{2.5a} \\
F_4 y &= 0. \tag{2.5b}
\end{aligned}
$$

Over $\mathcal{C}$, the observation o specifies the Gaussian density

$$\mathbf{p}(y,x) = k \exp -\frac{1}{2}(||(F_1 y - M_1 x)||^2 + ||F_3 y||^2) \tag{2.6}$$

where k is a constant which can be selected equal to 1, since we do not require that $\mathbf{p}$ should be normalized [1]. Thus, the specification of an observation of the form (2.1) is equivalent to the selection of a Gaussian density and constraint space pair $(\mathbf{p}(y,x)$, $\mathcal{C})$.

Proof: We only need to prove 1), since 2) is a direct consequence of 1). By performing a QR factorization of A, we can find an orthonormal matrix

$$T = \left[\begin{array}{c} K \\ L \end{array} \right] \tag{2.7a}$$

such that

$$TA = \left[\begin{array}{c} A_r \\ 0 \end{array} \right] , \tag{2.7b}$$

where A_r has full row rank. Denoting

$$TE = \left[\begin{array}{c} E_r \\ \overline{E} \end{array} \right] \tag{2.8}$$

and $\overline{B} = LB$, we see that

$$\overline{E} y = \overline{B} u . \tag{2.9}$$

Let

$$\overline{B} = U \left[\begin{array}{cc} 0 & \Sigma \\ 0 & 0 \end{array} \right] V , \tag{2.10}$$

with U and V orthonormal, and Σ diagonal positive definite, be a singular value decomposition of $\overline{B}$. Note that the block Σ has been placed in position (1,2) in order to ensure consistency with the choice (2.3) of input transformation. Then, the transformation

$$\left[\begin{array}{c} F_3 \\ F_4 \end{array} \right] = \left[\begin{array}{cc} \Sigma^{-1} & 0 \\ 0 & I \end{array} \right] U^T \overline{E} \qquad \left[\begin{array}{c} u_1 \\ u_3 \end{array} \right] = V u \tag{2.11}$$

yields the last two block rows of (2.4), where, because V is orthonormal, the covariance of Vu is identity. This implies u_1 and u_3 are independent with unit variance. Writing

$$KBV^T = \left[\begin{array}{cc} B_r & C_r \end{array} \right] , \tag{2.12}$$

---

[1] In other words, all densities which differ only by a multiplicative constant will be viewed as equivalent. This enables us to perform arbitrary linear transformations on the vectors x and y without keeping track of the Jacobians of the transformations.

and taking into account the third block row of (2.4), we also obtain

$$F_r y = A_r x + B_r u_1 \,, \tag{2.13a}$$

with

$$F_r \triangleq E_r - C_r F_3 \,. \tag{2.13b}$$

The assumption that B has full column rank implies B_r has full column rank. Consequently, by performing a QR decomposition of B_r, we can find an invertible matrix Z such that

$$Z B_r = \begin{bmatrix} I \\ 0 \end{bmatrix} \,. \tag{2.14}$$

Multiplying (2.13a) on the left by Z, and denoting

$$\begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = Z F_r \qquad M = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} = Z A_r \,, \tag{2.15}$$

then gives the first two block rows of (2.4), where the matrix M has full row rank since A_r has full row rank and Z is invertible. $\qquad\square$

Remarks:

1) An interesting feature of the above decomposition is that it relies exclusively on numerically stable operations, such as QR or singular value decompositions. In some sense, it can be viewed as just a formalization of procedures which occur repeatedly, but under different disguises, in the implementation of square-root Kalman filters [3].

2) Given an observation o of the form (2.1), the generation of the last two equations of (2.4) requires only the first step of the procedure employed to construct (2.4). Specifically, as indicated by (2.7a)–(2.7b), if L is a matrix whose rows form a basis of the left null space of A, so that $LA = 0$, the last two equations of (2.4) are given by

$$L E y = L B u \,. \tag{2.16}$$

Thus, the full decomposition procedure of Lemma 2.1 needs to be applied only if we seek to compute the whole canonical form (2.4).

An interesting feature of the constraints (2.5a) and (2.5b) defining $\mathcal{C}$ is that the second involves y only. The space

$$\mathcal{S} = \{ y \,:\, F_4 y = 0 \} \tag{2.17}$$

identifies the observations which are compatible with the model (2.1), so that the test $F_4 y = 0$ is actually the parity check that must be applied to an observation y to determine whether it can arise from the model (2.1). In this context, it is usually desirable to compress the row space of F_4, i.e to find an orthonormal matrix W such that

$$W F_4 = \begin{bmatrix} F_{4r} \\ 0 \end{bmatrix} \,, \tag{2.18}$$

where F_{4r} has full row rank. Then, the parity check test required to determine whether y belongs to $\mathcal{S}$ reduces to

$$F_{4r}y = 0 \,, \tag{2.19}$$

which involves only independent tests. Furthermore, if

$$\mathcal{X}(y) = \{\, x \ : \ M_2 x = F_2 y \text{ for } y \in \mathcal{S} \,\} \,, \tag{2.20}$$

the constraint space $\mathcal{C}$ can be expressed as

$$\mathcal{C} = \bigcup_{y \in \mathcal{S}} \left[\begin{array}{c} y \\ \mathcal{X}(y) \end{array} \right] \,, \tag{2.21}$$

i.e. given a valid observation $y \in \mathcal{S}$, the vector x is constrained to belong to the affine space $\mathcal{X}(y)$. This space describes the deterministic constraints to which x is subjected once y has been obtained and its validity established.

The affine space $\mathcal{X}(y)$ can be parametrized as follows. Let N_2 be a matrix whose columns form a basis of the right null space of M_2. Then, observing that

$$x_p(y) = M_2^T (M_2 M_2^T)^{-1} F_2 y \tag{2.22a}$$

is a particular solution of the equation

$$M_2 x = F_2 y \,, \tag{2.22b}$$

the vectors of $\mathcal{X}(y)$ admit the linear parametrization

$$x(\xi, y) = N_2 \xi + x_p(y) \tag{2.23}$$

where ξ is an arbitrary vector.

Thus, once the validity of the observation y has been established, the information available about x presents the following features:

(i) The density $\mathbf{p}(y, x)$ is concentrated on the affine space $\mathcal{X}(y)$, so that some components of x are known perfectly.

(ii) Some other components are completely unknown, which is reflected by the fact that the density $\mathbf{p}$ is uniformly distributed with respect to these variables.

(iii) The remaining components of x admit a Gaussian distribution.

We are therefore confronted with a situation where we have mixed information about the various components of x. Obviously, given a complete set of observations, we would like to find ourselves in a situation where x does not include any unknown component. This property can be described as follows.

Definition 2.1 *A vector x is said to be estimable if, given a linear parametrization $x(\xi, y)$ of $\mathcal{X}(y)$, the log-density $-\ln \mathbf{p}(y, x(\xi, y))$ is a nonsingular quadratic form of the vector ξ. Then, $\mathbf{p}(y, x)$ admits a unique maximum over $\mathcal{X}(y)$, and the integral*

$$\int g(x(\xi, y))\mathbf{p}(y, x(\xi, y))d\xi \tag{2.24}$$

exists for all functions $g(x)$ which grow at most polynomially as $\|x\| \to \infty$, so that all moments of the density $\mathbf{p}(y, x)$ exist over $\mathcal{X}(y)$.

Setting $g = 1$ in (2.24), we see that when x is estimable, the marginal density

$$\mathbf{p}(y) = \int \mathbf{p}(y, x(\xi, y))d\xi \tag{2.25}$$

can be defined over the space $\mathcal{S}$ of valid observations.

Lemma 2.2 *The vector x is estimable if and only if in the observation equation (2.1), A is one-to-one, or equivalently, iff in the canonical decomposition (2.2)–(2.4), the matrix*

$$M = \left[\begin{array}{c} M_1 \\ M_2 \end{array} \right] \tag{2.26}$$

is invertible.

Proof: Substituting the parametrization (2.23) of $\mathcal{X}(y)$ inside the density $\mathbf{p}(y, x)$ in (2.6) yields

$$-\ln \mathbf{p}(y, x(\xi, y)) = \frac{1}{2}(\|F_1 y - M_1 x_p(y) - M_1 N_2 \xi\|^2 + \|F_3 y\|^2), \tag{2.27}$$

which shows that $-\ln \mathbf{p}(y, x(\xi, y))$ is a nonsingular quadratic form of ξ if and only if $M_1 N_2$ is one to one, i.e. iff no vector of the null space of M_2 belongs also to the null space of M_1. This means that M must be invertible (note that by construction, M is onto). Since M is related to A through the transformation (2.2), this is equivalent to requiring that A should be one to one. $\qquad\square$

The expression (2.25) defines the marginal density $\mathbf{p}(y)$ only when x is estimable. However, given an observation of the form (2.1), we can always obtain an estimable parametrization by factoring A as

$$A = NM \tag{2.28}$$

with N one-to-one and M onto. Then, with $z = Mx$, the observation can be rewritten as

$$Ey = Nz + Bu, \tag{2.29}$$

where z is estimable. Thus, we can always find a parametrization which allows the evaluation of the marginal density $\mathbf{p}(y)$ through (2.25). In particular, in the coordinate system (2.2)–(2.4), we can select as estimable parameter vector

$$z = \left[\begin{array}{c} z_1 \\ z_2 \end{array} \right] = \left[\begin{array}{c} M_1 \\ M_2 \end{array} \right] x . \tag{2.30}$$

Integrating the density

$$\mathbf{p}(y, z) = \exp -\frac{1}{2}(||F_1 y - z_1||^2 + ||F_3 y||^2) \tag{2.31}$$

over the affine space $\mathcal{Z}(y) = \{ z : z_2 = F_2 y \}$ then gives

$$\mathbf{p}(y) = \exp -\frac{1}{2}||F_3 y||^2 \tag{2.32}$$

which is defined over $\mathcal{S}$.

A pair $(\mathbf{p}(y, x), \mathcal{C})$ can therefore be decomposed into two pairs:

(i) $(\mathbf{p}(y), \mathcal{S})$ which specifies the marginal density and constraint space for the measurement y. The density $\mathbf{p}(.)$ will be called the likelihood function of the observation o, and will be denoted as $L(o)$. Also, unless the context is clear, we shall use $\mathcal{S}(o)$ to denote the constraint space associated to o, so that $\mathcal{S}^{\perp}(o)$ will be the space of parity checks that must be applied to y to determine its validity.

(ii) The conditional density and constraint space pair $(\mathbf{p}(x|y) = \mathbf{p}(y, x)/\mathbf{p}(y), \mathcal{X}(y))$, where in the coordinate system (2.2)–(2.4), we have

$$\mathbf{p}(x|y) = \exp -\frac{1}{2}||F_1 y - M_1 x||^2 . \tag{2.33}$$

Thus, the major difference with observation models with nonsingular noise is that we must keep track of constraint spaces along with the usual densities or conditional densities.

B   ML and LS Estimation

The probabilistic set-up that we employ has the effect of blurring the usual distinction between Bayesian and ML estimation. On one hand, we model all prior information as observations. For example, to describe the measurement equation

$$y = Hx + Jw \tag{2.34}$$

with $x \sim \mathcal{N}(0, P)$ independent of $w \sim \mathcal{N}(0, I_m)$, we replace it by the equivalent model

$$\left[\begin{array}{c} 0 \\ I \end{array} \right] y = \left[\begin{array}{c} I \\ H \end{array} \right] x + \left[\begin{array}{cc} G & 0 \\ 0 & J \end{array} \right] \left[\begin{array}{c} v \\ w \end{array} \right] , \tag{2.35}$$

with $GG^T = P$ and the noise $v \sim \mathcal{N}(0, I_r)$ independent of w, where r denotes the rank of P. The only difference betwen (2.34) and (2.35) is that the prior information about x is modeled as an extra measurement. This gives the impression that we have adopted a pure ML point of view. Yet, unlike the usual ML philosophy, where $\mathbf{p}(y, x)$ is considered as a probability density for y with x unknown and fixed, $\mathbf{p}(y, x)$ is viewed here as a joint density for both y and x over the constraint space $\mathcal{C}$.

Then, the problem of estimating an estimable vector x from an observation o of the form (2.1) can be formulated as follows.

Definition 2.2 *Assume that x is estimable. Then the ML estimate $\hat{x}_{ML}(y)$ of x given $y \in \mathcal{S}$ is the argument of the maximum of $\mathbf{p}(x|y)$ (or equivalently, $\mathbf{p}(y, x)$) over $\mathcal{X}(y)$. The least-squares (LS) estimate $\hat{x}_{LS}(y)$ is the function $\hat{x}(y) \in \mathcal{X}(y)$ which minimizes*

$$E[||x - \hat{x}(y)||^2|y] \triangleq \int ||x(\xi, y) - \hat{x}(y)||^2 \mathbf{p}(x(\xi, y)|y)d\xi ,\tag{2.36}$$

where $x(\xi, y)$ is a linear parametrization of $\mathcal{X}(y)$.

It is well known, at least in the case of observations with nonsingular noise, that for linear Gaussian models, the ML and LS estimates are the same. This result remains true in the presence of deterministic information.

Theorem 2.1 *If x is estimable, its ML and LS estimates are identical, i.e.,*

$$\hat{x}_{ML}(y) = \hat{x}_{LS}(y)\tag{2.37a}$$

and

$$\mathbf{p}(y) = \mathbf{p}(y, \hat{x}_{ML}(y))\tag{2.37b}$$

for $y \in \mathcal{S}$, where the equality in (2.37b) holds modulo multiplication by a constant independent of y. Thus the likelihood function of o can be obtained either by computing the marginal density $\mathbf{p}(y)$ or by substituting the ML estimate $\hat{x}_{ML}(y)$ inside the joint density $\mathbf{p}(y, x)$.

Proof: We need only to modify slightly the usual derivation for Gaussian densities over unconstrained spaces. Without loss of generality, it can be assumed o has been brought to the canonical form (2.2)–(2.4), where since x is estimable, M is invertible. Denoting

$$M^{-1} = Q = \left[\begin{array}{cc} Q_1 & Q_2 \end{array} \right] ,\tag{2.38}$$

the space $\mathcal{X}(y)$ can be parametrized as

$$x(\xi, y) = Q_1\xi + Q_2 F_2 y ,\tag{2.39}$$

which yields

$$\mathbf{p}(x(\xi, y)|y) = \exp -\frac{1}{2}||F_1 y - \xi||^2 .\tag{2.40}$$

Its maximum is reached for $\hat{\xi}_{ML} = F_1 y$, so that

$$\hat{x}_{ML}(y) = Q \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} y \,. \tag{2.41}$$

Also, since $\mathbf{p}(\hat{x}_{ML}(y)|y) = 1$, (2.37b) holds.

Similarly, by employing the parametrization (2.39), we find

$$\int \|x(\xi, y) - \hat{x}(y)\|^2 \mathbf{p}(x(\xi, y))d\xi$$

$$= \int \|Q \begin{bmatrix} \xi \\ F_2 y \end{bmatrix} - \hat{x}(y)\|^2 \exp -\|F_1 y - \xi\|^2/2 \,,) \tag{2.42}$$

which is minimized by

$$\hat{x}_{LS}(y) = Q \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} y \,, \tag{2.43}$$

so that the ML and LS estimates coincide. $\qquad\square$

3 Primitives and Macros

We have now all the tools necessary to develop a probabilistic interpretation for the ML estimation primitives introduced in [7]. However, unlike [7], where the attention was focused exclusively on ML estimates and their associated error variances, we shall examine the effect of the primitives on all the subcomponents of an observation.

A Combinator

The first primitive we consider is the combination $o_1 \wedge o_2$ of two independent observations o_1 and o_2, where the symbol "$\wedge$" should be interpreted as the logical "and" connector.

Definition 3.1 *Consider two observations*

$$o_1 : \quad \begin{bmatrix} E_{11} & E_{1c} \end{bmatrix} \begin{bmatrix} y_1 \\ y_c \end{bmatrix} = \begin{bmatrix} A_{11} & A_{1c} \end{bmatrix} \begin{bmatrix} x_1 \\ x_c \end{bmatrix} + B_1 u_1 \tag{3.1a}$$

$$o_2 : \quad \begin{bmatrix} E_{2c} & E_{22} \end{bmatrix} \begin{bmatrix} y_c \\ y_2 \end{bmatrix} = \begin{bmatrix} A_{2c} & A_{22} \end{bmatrix} \begin{bmatrix} x_c \\ x_2 \end{bmatrix} + B_2 u_2 \,, \tag{3.1b}$$

where the noises u_1 and u_2 are independent with $u_i \sim \mathcal{N}(0, I)$ for $i = 1, 2$, and such that y_c and x_c denote respectively the measurement and parameter vector components which are common to both observations. The combined observation is given by

$$o_1 \wedge o_2 : \quad \begin{bmatrix} E_{11} & E_{1c} & 0 \\ 0 & E_{2c} & E_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_c \\ y_2 \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & A_{1c} & 0 \\ 0 & A_{2c} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_c \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} . \tag{3.2}$$

The combinator $\wedge$ is clearly commutative and associative. If

$$y = \begin{bmatrix} y_1 \\ y_c \\ y_2 \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_c \\ x_2 \end{bmatrix} \tag{3.3a}$$

$$\eta_1 = \begin{bmatrix} y_1 \\ y_c \end{bmatrix} \qquad \eta_2 = \begin{bmatrix} y_c \\ y_2 \end{bmatrix} \tag{3.3b}$$

$$\chi_1 = \begin{bmatrix} x_1 \\ x_c \end{bmatrix} \qquad \chi_2 = \begin{bmatrix} x_c \\ x_2 \end{bmatrix} , \tag{3.3c}$$

the constraint space $\mathcal{C}$ for $[\, y^T \; x^T \,]^T$ is obtained by taking the conjunction of the constraints defining $\mathcal{C}_1$ and $\mathcal{C}_2$, which concern respectively the vectors $[\, \eta_1^T \; \chi_1^T \,]^T$ and $[\, \eta_2^T \; \chi_2^T \,]^T$. Over $\mathcal{C}$, the density of $o_1 \wedge o_2$ is given by

$$\mathbf{p}(y, x) = \mathbf{p}_1(\eta_1, \chi_1)\mathbf{p}_2(\eta_2, \chi_2) , \tag{3.4}$$

where $\mathbf{p}_i(\eta_i, \chi_i)$ is the density over $\mathcal{C}_i$ associated to the observation o_i, with $i= 1, 2$.

B   Reduction and Coreduction

Next, consider an observation o of the form (2.1). It may contain information which is either redundant or irrelevant for the purpose of estimating x. This manifests itself by the existence of a sufficient statistic of smaller dimension than Ey. However, for our purposes we need a concept stronger than that of a sufficient statistic, which concerns only the compression of an observation vector. In addition, we need to ensure that the constraint space $\mathcal{X}(y)$ is not affected by the replacement of Ey by a sufficient statistic. This leads us to introduce the concept of *sufficient observation*, which is directly derived from the Neyman-Fisher factorization criterion [11], p. 39 for the sufficiency of a statistic.

Definition 3.2 *Given an observation o given by (2.1), we say that*

$$o_s \; : \quad KEy = KAx + KBu \tag{3.5}$$

is a sufficient observation for o if it has the same affine constraint space:

$$\mathcal{X}_s(y) = \mathcal{X}(y) , \tag{3.6}$$

and if, over this space,

$$\mathbf{p}_s(x|y) = \mathbf{p}(x|y) , \tag{3.7}$$

where $\mathbf{p}_s(x|y)$ and $\mathbf{p}(x|y)$ denote respectively the conditional densities (cf. (2.33)) corresponding to o_s and o.

If $\mathbf{p}(y,x)$ and $\mathbf{p}_s(y,x)$ are the densities of o and o_s, it is easy to verify that the condition (3.7) is equivalent to requiring

$$\mathbf{p}(y,x) = \mathbf{p}_s(y,x)\mathbf{q}(y) \tag{3.8}$$

where $\mathbf{q}(y)$ is independent of x, which except for the fact it must hold over the affine space (3.6), takes the usual form of the Neyman-Fisher factorization criterion. A simple situation where a sufficient observation can be identified by inspection is as follows.

Lemma 3.1 *Consider two observations*

$$o_0 : \quad E_0 y \;\; = \;\; A_0 x + B_0 u_0 \tag{3.9a}$$
$$o_1 : \quad E_1 y \;\; = \;\; B_1 u_1 \,, \tag{3.9b}$$

where u_0 and u_1 are two independent zero-mean Gaussian vectors with unit variance. Then, o_0 is a sufficient observation for $o = o_0 \wedge o_1$.

Proof: If $\mathbf{p}_0(y,x)$ and $\mathbf{p}_1(y)$ denote respectively the the densities of o_0 and o_1, the density of the joint observation o is given by

$$\mathbf{p}(y,x) = \mathbf{p}_0(y,x)\mathbf{p}_1(y) \,, \tag{3.10}$$

which is in the form (3.8). Also, since x does not appear in o_1, and the observation noises are independent,

$$\mathcal{X}(y) = \mathcal{X}_0(y) \,. \tag{3.11}$$

$\square$

Since all sufficient observations corresponding to a given observation need not have the same dimension, it is of interest to characterize those with the smallest size. The concept of minimal sufficient statistic [11], p. 41 can be extended to observations in the following manner.

Definition 3.3 *Consider an observation o of dimension q, i.e. such Ey has dimension q. It is minimal if it does not admit a sufficient observation o_s of lower dimension.*

Lemma 3.2 *An observation o is minimal if and only if the matrix A in (2.1) has full row rank, or equivalently, its likelihood function and parity space satisfy*

$$\mathbf{p}(y) = 1 \qquad \mathcal{S}^\perp(o) = \{0\} \,, \tag{3.12}$$

i.e. the observation vector is unconstrained and uniformly distributed.

Proof: Without loss of generality it can be assumed that o is in the form (2.2)–(2.4), in which case

$$A = \begin{bmatrix} M_1 \\ M_2 \\ 0 \\ 0 \end{bmatrix} \,. \tag{3.13}$$

From this representation, we can conclude A has full row rank iff its canonical decomposition does not include $F_3 y$ and $F_4 y$ components. According to the representations (2.32) and (2.17) for $\mathbf{p}(y)$ and $\mathcal{S}(o)$, this is equivalent to (3.12).

Conversely, suppose that A has full row rank, and that

$$o_s \quad : \quad KEy = KAx + KBu \tag{3.14}$$

is a sufficient statistic for o. Without loss of generality, we can assume that both o_s and o are in canonical form (this requires employing different representations of the noise vector u in o and o_s), in which case

$$KE = \begin{bmatrix} F_{s1} \\ F_{s2} \\ F_{s3} \\ F_{s4} \end{bmatrix} \qquad KA = \begin{bmatrix} M_{s1} \\ M_{s2} \\ 0 \\ 0 \end{bmatrix} \tag{3.15a}$$

$$E = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} \qquad A = \begin{bmatrix} M_1 \\ M_2 \end{bmatrix}, \tag{3.15b}$$

where since A has full row rank, the canonical decomposition of o does not include $F_3 y$ or $F_4 y$ components. Because o_s is a sufficient observation for o, it satisfies the conditions (3.6) and (3.7). Observing that since A has full row rank, the observations y are unconstrained, the condition (3.6) implies that the right null spaces of

$$\begin{bmatrix} F_{s2} & -M_{s2} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} F_2 & -M_2 \end{bmatrix} \tag{3.16}$$

are identical, so that there exists an invertible matrix U_2 such that

$$U_2 \begin{bmatrix} F_{s2} & -M_{s2} \end{bmatrix} = \begin{bmatrix} F_2 & -M_2 \end{bmatrix}. \tag{3.17}$$

Taking into account the expression (2.33) for conditional densities, the condition (3.7) implies

$$\| \begin{bmatrix} F_{s1} & -M_{s1} \end{bmatrix} \begin{bmatrix} y \\ x \end{bmatrix} \|^2 = \| \begin{bmatrix} F_1 & -M_1 \end{bmatrix} \begin{bmatrix} y \\ x \end{bmatrix} \|^2 + C(y) \tag{3.18}$$

for all vectors $[\, y^T \; x^T \,]^T$ belonging to the right null space of the matrices in (3.16), where $C(y)$ denotes an additive factor depending only on y. Consequently, there exist two matrices U_1 and U_{12}, with U_1 invertible, such that

$$\begin{bmatrix} U_1 & U_{12} \\ 0 & U_2 \end{bmatrix} \begin{bmatrix} F_{s1} & -M_{s1} \\ F_{s2} & -M_{s2} \end{bmatrix} = \begin{bmatrix} F_1 & -M_1 \\ F_2 & -M_2 \end{bmatrix}. \tag{3.19}$$

Since A is invertible, this implies that the observation o_s has dimension greater than or equal to o, so that o is minimal. The identity (3.19) indicates also that two minimal sufficient observations are necessarily related by a left multiplication with an invertible matrix K.

This is due to the fact that if o_s is minimal, its canonical decomposition (3.15a) does not include F_{s3} and F_{s4} blocks. □

To gain some intuition about the concept of minimal observation, consider a minimal observation of the form (2.1), where x is assumed estimable. Then A has both full row and column rank, so that it is invertible, and according to (2.41), $\hat{x}_{ML} = A^{-1}Ey$ is the ML estimate of x. Without loss of information, we can premultiply (2.1) by A^{-1}, which yields

$$\hat{x}_{ML} = x + Gu \qquad (3.20)$$

with $G = A^{-1}B$, so that $P = GG^T$ is the error covariance of the ML estimate. Thus, in this case, a minimal observation represents just a way of coding the ML estimate and error covariance of x without including any extraneous information. More generally, when x is not estimable, a minimal observation represents just a code for the ML estimate and error variance of its estimable part $z = Ax$.

The identities (3.12) show that once a minimal sufficient observation has been constructed, it contains no useful information about the constraint space $\mathcal{S}(o)$ and likelihood function $L(o)$ of the original observation o. The *reduction primitive* described below provides a mechanism for constructing minimal sufficient observations, while at the same time keeping track of the constraint space and likelihood function of o.

Definition 3.4 (reduction and coreduction) *Given an observation o in the canonical form (2.2)–(2.4), the reduced observation $R\{o\}$ and its complement, the coreduction $\overline{R}\{o\}$, are given by*

$$R\{o\} \; : \quad \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} y \;\; = \;\; \begin{bmatrix} M_1 \\ M_2 \end{bmatrix} x + \begin{bmatrix} u_1 \\ 0 \end{bmatrix} \qquad (3.21)$$

$$\overline{R}\{o\} \; : \quad \begin{bmatrix} F_3 \\ F_4 \end{bmatrix} y \;\; = \;\; \begin{bmatrix} u_3 \\ 0 \end{bmatrix} . \qquad (3.22)$$

Combining Lemmas 2.1 and 3.2, we obtain the following result, which justifies our choice of reduction primitive.

Theorem 3.1

1) $R\{o\}$ is a minimal sufficient observation for o, and

2) the constraint space $\mathcal{S}(o)$ and likelihood function $L(o)$ satisfy

$$\mathcal{S}(o) = \mathcal{S}(\overline{R}\{o\}) \qquad L(o) = L(\overline{R}\{o\}) . \qquad (3.23)$$

Thus, the reduction operation completely decouples the ML estimation of x from the problem of computing the constraint space and likelihood function of o: the minimal observation $R\{o\}$ contains all the information necessary for the ML estimation of x, whereas the

coreduction $\overline{R}\{o\}$ contains all the information concerning the constraint space and likelihood function of o.

Finally, note that although the primitives $R\{.\}$ and $\overline{R}\{.\}$ are represented by different symbols, because they can both be implemented with the decomposition algorithm of Lemma 2.1, it can be argued they form only one primitive. However, from a computational point of view, it is preferable to implement $\overline{R}\{.\}$ and $R\{.\}$ as distinct entities. Indeed, as indicated by the Remark 2) following the proof of Lemma 2.1, given an observation o of the form (2.1), the generation of $\overline{R}\{o\}$ requires only the first step of the procedure employed to find $R\{o\}$. Specifically, if L is a matrix whose rows form a basis of the left null space of A, the coreduction $\overline{R}\{o\}$ takes the form (2.16). Thus, the full decomposition procedure of Lemma 2.1 needs to be applied only if we seek to compute *both* $R\{o\}$ and $\overline{R}\{o\}$.

C   Mutation and Other Macros

In addition to the three primitives: $\wedge$, $R\{.\}$, $\overline{R}\{.\}$ previously introduced, we need to introduce a fourth primitive whose effect is to change the status of a vector inside an observation from known to unknown, or vice-versa. This primitive will be useful to formulate the concept of conditional observation, where certain vectors need to be treated temporarily as known, even though they are actually unknown.

Definition 3.5 *Given an observation o of the form*

$$o : \quad Ey = \begin{bmatrix} A_0 & A_1 \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} + Bu , \tag{3.24}$$

the observation

$$M_{x_1}\{o\} : \quad \begin{bmatrix} E & -A_1 \end{bmatrix} \begin{bmatrix} y \\ x_1 \end{bmatrix} = A_0 x_0 + Bu \tag{3.25}$$

obtained by changing the status of the vector x_1 from unknown to known, is said to be a mutation of x_1 inside o. Similarly, given

$$o' : \quad \begin{bmatrix} E_0 & E_1 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = Ax + Bu , \tag{3.26}$$

the observation

$$M_{y_1}\{o'\} : \quad E_0 y_0 = \begin{bmatrix} A & -E_1 \end{bmatrix} \begin{bmatrix} x \\ y_1 \end{bmatrix} + Bu , \tag{3.27}$$

generated by changing the status of y_1 from known to unknown constitutes a mutation of y_1 inside o'.

Clearly, two consecutive mutations of the same vector yield the original observation. Consider now an observation of the form (3.24), and suppose we are no longer interested in the vector x_0 and just seek to "extract" the information about y and x_1 contained in (3.24). This amounts to computing the marginal density $\mathbf{p}(y, x_1)$ and constraint space $\mathcal{C}_1$

for $[\, y^T \; x_1^T \,]^T$ from the pair $(\mathbf{p}(y, x_0, x_1), \mathcal{C})$ associated to o. Referring to the property (3.23) of the coreduction $\overline{R}\{.\}$, we see it provides a mechanism for extracting the marginal density and constraint space associated to the measurement vector y. If we want to include the vector x_1 in the marginal computation, we need only to change its status and view it as a known vector before applying the $\overline{R}\{.\}$ primitive. After $\overline{R}\{.\}$ has been applied, we can of course convert back x_1 to its original "unknown" status. This motivates the construction of the following two macros, which are expressed entirely in terms of the four primitives already introduced.

Definition 3.6 (conditional observation, extraction) *Given an observation of the form (3.24), the conditional observation of o given x_1, and the extraction of x_1 from o, are defined respectively as*

$$o \mid x_1 \;\; = \;\; R \circ M_{x_1}\{o\} \tag{3.28a}$$

$$X_{x_1}\{o\} \;\; = \;\; M_{x_1} \circ \overline{R} \circ M_{x_1}\{o\}\,, \tag{3.28b}$$

where "$\circ$" denotes the composition of maps.

Taking into account the comment preceding (2.16), we see that if L is a matrix whose rows form a basis of the left null space of A_0, $X_{x_1}\{o\}$ can be expressed as

$$X_{x_1}\{o\} \;\; : \;\;\;\; LEy = LA_1 x_1 + LBu\,. \tag{3.29}$$

Thus, in the context of statistical models containing mixed deterministic/stochastic information, the macros $X_{x_1}\{o\}$ and $o \mid x_1$ perform respectively the computation of marginal and conditional densities and constraint spaces.

D   Properties of the primitives and macros

This subsection describes several useful properties of the primitives and macros we have introduced. We begin with simple formulas involving a single observation. Their proof is immediate.

Lemma 3.3 *The following formulas hold:*

$$o \;\; = \;\; R\{o\} \wedge \overline{R}\{o\} \tag{3.30a}$$

$$o \;\; = \;\; X_{x_1}\{o\} \wedge M_{x_1}\{o \mid x_1\}\,, \tag{3.30b}$$

where in (3.30b), we assume o has the form (3.24).

The extracted observation $X_{x_1}\{o\}$ has the following properties.

Theorem 3.2

1) *Given an observation o of the form (3.24), the ML estimates of the estimable part of x_1 based on o and $X_{x_1}\{o\}$ are the same.*

*2) The extraction operation preserves the information contained in o concerning the mea-
surement vector y, in the sense that*

$$\overline{R} \circ X_{x_1}\{o\} = \overline{R}\{o\} \ . \tag{3.31}$$

In particular, we have

$$\mathcal{S}(o) = \mathcal{S}(X_{x_1}\{o\}) \qquad L(o) = L(X_{x_1}\{o\}) \ . \tag{3.32}$$

Proof: Consider the decomposition (3.30b) of o. Since $o \mid x_1$ is by definition a minimal observation, according to the Lemma 3.2, its marginal density with respect to the pair (y, x_1) is a constant, and $o \mid x_1$ does not impose any constraint on (y, x_1). But applying the mutation operation M_{x_1} does not affect either the joint density of y and x_1 or their constraint space. Thus, in the decomposition $X_{x_1}\{o\} \wedge M_{x_1}\{o \mid x_1\}$ of observation o, the second component does not provide any information about x_1. This proves the first statement of of the theorem.

Also, since $\overline{R}\{o\}$ just encodes the constraint space and likelihood function of y, when applying $\overline{R}$ to the decomposition $X_{x_1}\{o\} \wedge M_{x_1}\{o \mid x_1\}$, we can ignore the second observation. This proves formula (3.31). Then, (3.32) is just a consequence of (3.31) and (3.23).   □

We now move to properties involving two observations. Note first that for an arbitrary vector z, we have the obvious identity

$$M_z\{o_1 \wedge o_2\} = M_z\{o_1\} \wedge M_z\{o_2\} \ . \tag{3.33}$$

Next, observe that the presence of the coupling parameter vector x_c in expression (3.2) has the effect of complicating the properties of the combinator $\wedge$. Thus, it is of interest introduce the absence of coupling vector x_c as a special property. On the other hand, the presence of a coupling measurement y_c causes no particular trouble, as will be seen below.

Definition 3.7 (independent observations) *Two observations o_1 and o_2 with the structure (3.1a)–(3.1b) are said to be independent, which is denoted as*

$$o_1 \perp o_2 \ ,$$

if they do not include any unknown coupling vector x_c, or, more generally, if, for $i = 1$ or 2, the extracted observation $X_{x_c}\{o_i\}$ does not depend on x_c, i.e.

$$X_{x_c}\{o_i\} = \overline{R}\{o_i\} \ . \tag{3.34}$$

The property (3.34) is apparently weaker than the absence of coupling, since it only requires that o_i for $i = 1$ or 2 should convey no information about the coupling vector x_c. However, as the following lemma indicates, when (3.34) is satisfied, the observations o_1 and o_2 can be reparametrized so as to make them decoupled.

Lemma 3.4 *If an observation o_1 of the form (3.1a) satisfies (3.34), there exists a matrix T such that o_1 can be rewritten in terms of the new vector*

$$x'_1 = x_1 + Tx_c \, , \qquad (3.35a)$$

as

$$o_1 : \quad \begin{bmatrix} E_{11} & E_{1c} \end{bmatrix} \begin{bmatrix} y_1 \\ y_c \end{bmatrix} = A_{11} x'_1 + B_1 u_1 \, , \qquad (3.35b)$$

which is now decoupled from the observation o_2 of (3.1b).

Proof: If $X_{x_c}\{o_1\}$ is independent of x_c, the expression (3.29) for the extraction operation indicates that if L is a basis of the left null space of A_{11}, we must have

$$LA_{1c} = 0 \qquad (3.36)$$

i.e. L must be in the left null space of A_{1c}. Consequently, the column space of A_{1c} is contained in the column space of A_{11}, so that we can find a matrix T such that

$$A_{1c} = A_{11} T \, . \qquad (3.37)$$

Substituting this relation inside (3.1a), and defining x'_1 as in (3.35a), gives (3.35b). □

In general, because of the presence of the coupling vector x_c, the reduction $R\{o_1 \wedge o_2\}$, likelihood function $L(o_1 \wedge o_2)$ and constraint space $\mathcal{S}(o_1 \wedge o_2)$ cannot be expressed purely in terms of $(R\{o_i\}, L(o_i), \mathcal{S}(o_i))$ with $i = 1, 2$. However, when o_1 and o_2 are independent, we have the following expressions.

Lemma 3.5 *If $o_1 \perp o_2$, the combined observation $o = o_1 \wedge o_2$ satisfies*

$$R\{o_1 \wedge o_2\} = R\{o_1\} \wedge R\{o_2\} \qquad (3.38a)$$
$$\overline{R}\{o_1 \wedge o_2\} = \overline{R}\{o_1\} \wedge \overline{R}\{o_2\} \, . \qquad (3.38b)$$

In particular, we have

$$\mathcal{S}^\perp(o) = \mathcal{S}^\perp(o_1) + \mathcal{S}^\perp(o_2) \qquad (3.39a)$$
$$L(o) = L(o_1)L(o_2) \, , \qquad (3.39b)$$

where $+$ denotes the (not necessarily direct) sum of vector spaces. Furthermore, if there is no coupling measurement y_c, the sum in (3.39a) becomes direct.

Proof: Select a coordinate system such that o_1 and o_2 have no coupling vector x_c. Next, observe that the canonical decomposition of an observation given in Lemma 2.1 depends only on the matrix pair (A, B). Since the noises u_1 and u_2 are independent and o_1 and o_2 have no coupling vector, the canonical decomposition (2.2)–(2.4) for $o_1 \wedge o_2$ can be obtained by reducing and coreducing *separately* o_1 and o_2, and then forming $R\{o_1\} \wedge R\{o_2\}$ and

$\overline{R}\{o_1\} \wedge \overline{R}\{o_2\}$, which justifies (3.38a)–(3.38b). Then, using the representations (2.17) and (2.32) of $\mathcal{S}(o)$ and $L(o)$, gives (3.39a)–(3.39b).

The property (3.39a), which is expressed in terms of the parity spaces of o_1 and o_2, can be written equivalently in terms of their constraint spaces as

$$\mathcal{S}(o) = \mathcal{S}(o_1) \cap \mathcal{S}(o_2) . \tag{3.40}$$

Furthermore, when the observations o_1 and o_2 do not contain a common measurement component y_c, the parity check tests for o_1 and o_2 concern different measurement vectors y_1 and y_2, so that the sum in (3.39a) can be replaced by a direct sum. $\square$

As a corollary of the above lemma, we have the following result.

Corollary 3.1 *If $o_1 \perp o_2$, then for any unknown vector z,*

$$X_z\{o_1 \wedge o_2\} = X_z\{o_1\} \wedge X_z\{o_2\} . \tag{3.41}$$

E   Recursion Principle

We have now at our disposal all the tools necessary to develop recursive algorithms for computing ML estimates, and to evaluate the constraint set and likelihood function associated to a stream of observations. The algorithms that we shall derive rely on the recursion scheme described below, which is a consequence of Theorems 3.1 and 3.2.

Theorem 3.3 *Consider the observations*

$$o_1 : \qquad E_{11}y_1 + E_{1c}y_c \quad = \quad A_{11}x_1 + A_{1c}x_c + B_1 u_1 \tag{3.42a}$$

$$o_2 : \qquad E_{2c}y_c + E_{22}y_2 \quad = \quad A_{2c}x_c + A_{22}x_2 + B_2 u_2 , \tag{3.42b}$$

where u_1 and u_2 are two zero-mean independent Gaussian vectors with unit covariance. Then

$$R_\circ X_{x_2}\{o_1 \wedge o_2\} = R_\circ X_{x_2}\{R_\circ X_{x_c}\{o_1\} \wedge o_2\} \tag{3.43}$$

and

$$\overline{R}\{o_1 \wedge o_2\} = \overline{R}_\circ X_{x_2}\{R_\circ X_{x_c}\{o_1\} \wedge o_2\} \ \wedge \ \overline{R}_\circ X_{x_c}\{o_1\} , \tag{3.44}$$

which in turn implies

$$\mathcal{S}^\perp(o_1 \wedge o_2) \quad = \quad \mathcal{S}^\perp(\overline{R}_\circ X_{x_2}\{R_\circ X_{x_c}\{o_1\} \wedge o_2\}) + \mathcal{S}^\perp(\overline{R}_\circ X_{x_c}\{o_1\}) \tag{3.45a}$$

$$L(o_1 \wedge o_2) \quad = \quad L(\overline{R}_\circ X_{x_2}\{R_\circ X_{x_c}\{o_1\} \wedge o_2\})L(\overline{R}_\circ X_{x_c}\{o_1\}) , \tag{3.45b}$$

where the sum in (3.45a) can be replaced by a direct sum whenever o_1 and o_2 do not contain a common measurement y_c.

Proof: We first derive (3.43). Decomposing o_1 as in (3.30a), and observing that $R\{o_1\} \wedge o_2$ and $\overline{R}\{o_1\}$ are independent, the property (3.41) gives

$$X_{x_2}\{o_1 \wedge o_2\} = X_{x_2}\{R\{o_1\} \wedge o_2\} \wedge X_{x_2} \circ \overline{R}\{o_1\} . \tag{3.46}$$

Since x_2 does not appear in o_1, $X_{x_2} \circ \overline{R}\{o_1\} = \overline{R}\{o_1\}$. We can thus apply (3.38a), which yields

$$
\begin{aligned}
R\{X_{x_2}\{R\{o_1\} \wedge o_2\} \wedge \overline{R}\{o_1\}\} &= R \circ X_{x_2}\{R\{o_1\} \wedge o_2\} \wedge R \circ \overline{R}\{o_1\} \\
&= R \circ X_{x_2}\{R\{o_1\} \wedge o_2\} .
\end{aligned}
\tag{3.47}
$$

Now, according to (3.30b), o_1 can be decomposed as

$$o_1 = X_{x_c}\{o_1\} \wedge M_{x_c}\{o_1 \mid x_c\} , \tag{3.48}$$

where $X_{x_c}\{o_1\}$ and $M_{x_c}\{o_1 \mid x_c\}$ are independent since $M_{x_c}\{o_1 \mid x_c\}$ does not convey any information about x_c. Applying the reduction operation R to both sides, and noting that $R \circ M_{x_c}\{o_1 \mid x_c\}$ is independent from observation $R \circ X_{x_c}\{o_1\} \wedge o_2$, we can apply again (3.41). This gives

$$
\begin{aligned}
X_{x_2}\{(R \circ X_{x_c}\{o_1\} \wedge o_2) \wedge R \circ M_{x_c}\{o_1 | x_c\}\} &= X_{x_2}\{(R \circ X_{x_c}\{o_1\} \wedge o_2)\} \wedge X_{x_2} \circ R \circ M_{x_c}\{o_1 | x_c\} \\
&= X_{x_2}\{(R \circ X_{x_c}\{o_1\} \wedge o_2)\} .
\end{aligned}
\tag{3.49}
$$

Then, combining (3.47) and (3.49), we get (3.43).

Next, we prove (3.44). Using (3.46) and then (3.49) yields

$$
\begin{aligned}
X_{x_2}\{o_1 \wedge o_2\} &= X_{x_2}\{R\{o_1\} \wedge o_2\} \wedge X_{x_2} \circ \overline{R}\{o_1\} \\
&= X_{x_2}\{R\{o_1\} \wedge o_2\} \wedge \overline{R}\{o_1\} \\
&= X_{x_2}\{R \circ X_{x_c}\{o_1\} \wedge o_2 \wedge R \circ M_{x_c}\{o_1 \mid x_c\}\} \wedge \overline{R}\{o_1\} \\
&= X_{x_2}\{R \circ X_{x_c}\{o_1\} \wedge o_2\} \wedge \overline{R}\{o_1\} .
\end{aligned}
\tag{3.50}
$$

But

$$\overline{R} \circ X_{x_2}\{o_1 \wedge o_2\} = \overline{R}\{o_1 \wedge o_2\} , \tag{3.51}$$

and since observations $X_{x_2}\{R \circ X_{x_c}\{o_1\} \wedge o_2\}$ and $\overline{R}\{o_1\}$ are independent, by using (3.38b), we find

$$
\begin{aligned}
\overline{R}\left\{X_{x_2}\{R \circ X_{x_c}\{o_1\} \wedge o_2\} \wedge \overline{R}\{o_1\}\right\} &= \overline{R} \circ X_{x_2}\{R \circ X_{x_c}\{o_1\} \wedge o_2\} \wedge \overline{R}\{o_1\} \\
&= \overline{R} \circ X_{x_2}\{R \circ X_{x_c}\{o_1\} \wedge o_2\} \wedge \overline{R} \circ X_{x_c}\{o_1\} ,
\end{aligned}
\tag{3.52}
$$

where (3.52) uses the identity $\overline{R} \circ X_{x_c} = \overline{R}$. Combining (3.50)–(3.52) gives (3.44). $\qquad\square$

Thus our approach consists of estimating vectors in the order in which they arise in observations, and then discard these vectors for the purpose of processing subsequent observations where they do not appear. The constraint space $\mathcal{S}(o)$ is obtained by juxtaposition of the old constraints with new constraints for the latest observations. $L(o)$ is obtained by multiplying the old likelihood function with a term reflecting the new information, i.e. the "innovation", contained in the most recent measurement.

4 Filtering and Smoothing of Descriptor Systems and Application to Failure Detection

To illustrate the primitives and other macros introduced in the previous section, we solve the filtering and smoothing problems for linear descriptor systems, and then apply our results to a failure detection problem. Linear descriptor systems admit dynamics of the form

$$d_{k+1} : \qquad 0 = -E_{k+1}x_{k+1} + A_k x_k + B_k u_k \,, \tag{4.1}$$

with observations

$$m_{k+1} : \qquad y_{k+1} = C_{k+1}x_{k+1} + D_k u_k \tag{4.2}$$

for $0 \leq k \leq N-1$, where u_k is a white Gaussian noise such that $u_k \sim \mathcal{N}(0, I)$. The case $E_{k+1} = I$ corresponds to the usual state-space models, but E_{k+1} need not be invertible in general. In fact, the matrices E_{k+1} and A_k are not even required to be square, and the size of the state x_k can change with k. The presence of deterministic information is reflected here by the fact that the dynamics (4.1) may contain fixed algebraic constraints between entries of x_{k+1} and x_k, and the covariance $D_k D_k^T$ of the observation noise may not be invertible, thus resulting in a singular estimation problem.

The filtering problem for descriptor systems has been examined by several researchers [12], [13], [14], [7]. For the case when x_k is estimable from past observations, explicit recursions for the optimum filter and its associated Riccati equation are given in [13], which contains also an analysis of the stady-state convergence of the optimum filter. Square-root implementations of the descriptor Kalman filter were proposed in [15] and [14]. The case where x_k is not estimable from past observations (but may be estimable from past and future observations combined) is considered in [7], where the filtering and smoothing formulas were expressed for the first time in terms of the reduction and extraction primitives employed here. However, because a probabilistic interpretation of the primitives was not yet available, the solution of the filtering problem presented in [7] does not include the generation of an innovations process, or parity checks for evaluating the model validity in function of the received observations. Both of these items have important practical applications, since as will be shown below the computation of the likelihood function associated to model (4.1)–(4.2) relies on the innovations process. Furthermore, a wide class of failure detection algorithms [16], [17] relies on parity check tests of the type discussed here.

When considering the system (4.1)–(4.2), we make no distinction between dynamics and measurements, and view all linear relations relating different random variables as observations for these variables. From this point of view, it is convenient to combine the dynamics and measurement equations (4.1)–(4.2) into a single observation $o_{k+1} = d_{k+1} \wedge m_{k+1}$ of the form

$$o_{k+1} : \qquad z_{k+1} = \begin{bmatrix} 0 \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} -E_{k+1} \\ C_{k+1} \end{bmatrix} x_{k+1} + \begin{bmatrix} A_k \\ 0 \end{bmatrix} x_k + \begin{bmatrix} B_k \\ D_k \end{bmatrix} u_k \tag{4.3}$$

for $0 \le k \le N - 1$. The information available about the initial and final states is assumed to have the separable form

$$o_0 : \quad z_0 \;=\; -E_0 x_0 + B_{-1} u_{-1} \tag{4.4a}$$

$$o_{N+1} : \quad z_{N+1} \;=\; A_N x_N + B_N u_N \,. \tag{4.4b}$$

A Filtering

Let

$$P_k = \bigwedge_{j=0}^{k} o_j \tag{4.5}$$

be the observation obtained by combining all past observations up to time k. Then, the filtered observation describing the information about x_k contained in P_k, the constraint space of P_k, and its likelihood function, are denoted respectively by

$$\hat{o}^f_{k|k} \;=\; R \circ X_{x_k} \{P_k\} \tag{4.6a}$$

$$\mathcal{S}_k = \mathcal{S}\{P_k\} \qquad L_k = L(P_k) \,. \tag{4.6b}$$

Applying the recursion principle of Theorem 3.3, we obtain the *Kalman filtering recursions*

$$\hat{o}^f_{k+1|k+1} \;=\; R \circ X_{x_{k+1}} \{\hat{o}^f_{k|k} \wedge o_{k+1}\} \tag{4.7a}$$

$$\mathcal{S}^{\perp}_{k+1} \;=\; \mathcal{S}^{\perp}(\overline{R} \circ X_{x_{k+1}} \{\hat{o}^f_{k|k} \wedge o_{k+1}\}) \oplus \mathcal{S}^{\perp}_k \tag{4.7b}$$

$$L_{k+1} \;=\; L(\overline{R} \circ X_{x_{k+1}} \{\hat{o}^f_{k|k} \wedge o_{k+1}\}) L_k \tag{4.7c}$$

with initial conditions

$$\hat{o}^f_{0|0} = R\{o_0\} \qquad \mathcal{S}_0 = \mathcal{S}(\overline{R}\{o_0\}) \qquad L_0 = L(\overline{R}\{o_0\}) \,, \tag{4.8}$$

where $\oplus$ denotes the direct sum of vector spaces. These recursions, which obviously constitute a rather unusual form of the Kalman filter, are totally expressed in terms of the primitives and macros introduced in Section 3. To interpret the expressions obtained for $\mathcal{S}^{\perp}_k$ and L_k, note from the definition (3.22) of $\overline{R}\{.\}$ that the observation $\overline{R} \circ X_{x_{k+1}} \{\hat{o}^f_{k|k} \wedge o_{k+1}\}$ can be decomposed as

$$\begin{bmatrix} \nu_{k+1} \\ \pi_{k+1} \end{bmatrix} = \begin{bmatrix} w_{k+1} \\ 0 \end{bmatrix} \,, \tag{4.9}$$

where $w_{k+1} \sim \mathcal{N}(0, I)$. The vectors ν_{k+1} and π_{k+1} correspond respectively to the innovations and parity checks of the given model, and with respect to these vectors, the recursions (4.7b)–(4.7c) take the form

$$\mathcal{S}^{\perp}_{k+1} \;=\; \{y_j \,, 0 \le j \le k+1 \,:\, \pi_{k+1} = 0\}^{\perp} \oplus \mathcal{S}^{\perp}_k \tag{4.10a}$$

$$L_{k+1} \;=\; \exp{-\frac{1}{2}}\|\nu_{k+1}\|^2 L_k \tag{4.10b}$$

so that, as for nonsingular filtering problems, the log-likelihood function $l_k = -\ln L_k$ is obtained by summing the squared innovations.

Interpretation: The algorithm (4.7a)–(4.7c) is quite compact and constitutes by itself a complete implementation of the Kalman filter. However, for interpretation purposes, it can be transcribed into a form closer to the usual Kalman filter if we assume that x_k is estimable from P_k, so that $[\,-E_k^T\ C_k^T\,]^T$ has full rank for all k. In this case, $\hat{o}_{k|k}^f$ can be written as

$$\hat{o}_{k|k}^f\ :\qquad \hat{x}_{k|k}^f = x_k + \Sigma_{k|k}^{f\,1/2}v_k\ ,\tag{4.11}$$

where $\hat{x}_{k|k}^f$ and $\Sigma_{k|k}^f$ denote respectively the ML estimate and error covariance matrix of x_k given P_k, and $v_k \sim \mathcal{N}(0, I)$ is independent of the model noise u_k. Also, given an arbitary symmetric nonnegative definite matrix Σ, $\Sigma^{1/2}$ denotes one of its square-roots. Then the Kalman filtering algorithm (4.7a)–(4.7c) can be broken into two steps.

(i) Time update: We consider $\hat{o}_{k|k}^f \wedge d_{k+1}$, which takes the form

$$\begin{bmatrix} \hat{x}_{k|k}^f \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -E_{k+1} \end{bmatrix} x_{k+1} + \begin{bmatrix} I \\ A_k \end{bmatrix} x_k + \begin{bmatrix} \Sigma_{k|k}^{f\,1/2} & 0 \\ 0 & B_k \end{bmatrix}\begin{bmatrix} v_k \\ u_k \end{bmatrix}\ .\tag{4.12}$$

The predicted observation $\hat{o}_{k+1|k}^f \overset{\triangle}{=} X_{x_{k+1}}\{\hat{o}_{k|k}^f \wedge d_{k+1}\}$ is obtained by premultiplying (4.12) by $[\,-A_k\ I\,]$, which yields

$$\hat{o}_{k+1|k}^f\ :\qquad -A_k\hat{x}_{k|k}^f = -E_{k+1}x_{k+1} + B_k u_k - A_k\Sigma_{k|k}^{f\,1/2}v_k\ .\tag{4.13}$$

Since extractions do not change observation constraint spaces or likelihood functions, we have

$$\mathcal{S}_{k+1|k}\ \overset{\triangle}{=}\ \mathcal{S}(\hat{o}_{k+1|k}^f) = \mathcal{S}_k \tag{4.14a}$$

$$L_{k+1|k}\ \overset{\triangle}{=}\ L(\hat{o}_{k+1|k}^f) = L_k\ .\tag{4.14b}$$

(ii) Measurement update: Consider now the observation $\hat{o}_{k+1|k}^f \wedge m_{k+1}$:

$$\begin{bmatrix} -A_k\hat{x}_{k|k}^f \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} -E_{k+1} \\ C_{k+1} \end{bmatrix} x_{k+1} + \begin{bmatrix} B_k & -A_k\Sigma_{k|k}^{f\,1/2} \\ D_k & 0 \end{bmatrix}\begin{bmatrix} u_k \\ v_k \end{bmatrix}\ .\tag{4.15}$$

The transformation matrix S_{k+1} needed to bring this observation to the canonical form (2.2)–(2.4) can be partitioned as

$$S_{k+1} = \begin{bmatrix} S_{k+1\,1} \\ S_{k+1\,2} \\ S_{k+1\,3} \\ S_{k+1\,4} \end{bmatrix}\ .\tag{4.16}$$

With this notation, the reduced observation $\hat{o}^f_{k+1|k+1} = R\{\hat{o}^f_{k+1|k} \wedge m_{k+1}\}$ takes the form

$$\begin{bmatrix} S_{k+1\,1} \\ S_{k+2\,2} \end{bmatrix} \begin{bmatrix} -A_k \hat{x}^f_{k|k} \\ y_{k+1} \end{bmatrix} = M_{k+1} x_{k+1} + \begin{bmatrix} v_{k+1} \\ 0 \end{bmatrix} , \tag{4.17}$$

with $v_{k+1} \sim \mathcal{N}(0, I)$. Because of the estimability of x_{k+1}, M_{k+1} is invertible, and upon left multiplication of (4.17) by

$$M^{-1}_{k+1} = \begin{bmatrix} Q_{k+1\,1} & Q_{k+1\,2} \end{bmatrix} , \tag{4.18}$$

we obtain

$$\hat{x}^f_{k+1|k+1} : \qquad \hat{x}^f_{k+1|k+1} = x_{k+1} + \Sigma^{f\,1/2}_{k+1|k+1} v_{k+1} \tag{4.19}$$

with

$$\hat{x}^f_{k+1|k+1} = M^{-1}_{k+1} \begin{bmatrix} S_{k+1\,1} \\ S_{k+1\,2} \end{bmatrix} \begin{bmatrix} -A_k \hat{x}^f_{k|k} \\ y_{k+1} \end{bmatrix} \tag{4.20a}$$

$$\Sigma^{f\,1/2}_{k+1|k+1} = Q_{k+1\,1} . \tag{4.20b}$$

Similarly, $\overline{R}\{\hat{o}^f_{k+1|k} \wedge o_{k+1}\}$ is given by

$$\begin{bmatrix} S_{k+1\,3} \\ S_{k+1\,4} \end{bmatrix} \begin{bmatrix} -A_k \hat{x}^f_{k|k} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} \nu_{k+1} \\ \pi_{k+1} \end{bmatrix} = \begin{bmatrix} w_{k+1} \\ 0 \end{bmatrix} , \tag{4.21}$$

so that the innovations ν_{k+1} and parity check vector π_{k+1} are linear combinations of $\hat{x}^f_{k|k}$ and y_{k+1}.

Note that the above discussion merely provides a step by step account of the operations appearing in the Kalman filtering algorithm (4.7a)–(4.7c). However our goal is ultimately to hide all implementation details of this type by focusing on the larger picture provided by the primitives and macros we have introduced.

B Two-filter Smoother

To discuss two-filter implementations of the smoother associated to the model (4.1)–(4.2), it is convenient to construct the backward Kalman filter for this model. Thus, let

$$F_k = \bigwedge_{j=k}^{N+1} o_j \tag{4.22}$$

be the combination of all observations in the future of k. Then the backward predicted observation $\hat{o}^b_{k|k+1}$ describing the information about x_k contained in F_{k+1} is given by

$$\hat{o}^b_{k|k+1} = R \circ X_{x_k} \{F_{k+1}\} . \tag{4.23}$$

According to Theorem 3.2, it satisfies the backward Kalman recursion (in predicted form)

$$\hat{o}^b_{k-1|k} = R \circ X_{x_{k-1}} \{ o^b_{k|k+1} \wedge o_k \} \tag{4.24a}$$

$$\hat{o}^b_{N+1|N+2} = R \{ o_{N+1} \} . \tag{4.24b}$$

Then, if

$$O = \bigwedge_{j=0}^{N+1} o_j \tag{4.25}$$

represents the combination of all observations, the smoothed observation

$$\hat{o}^s_k = R \circ X_{x_k} \{ O \} \tag{4.26}$$

describes the information about x_k contained in all observations. Instead of having to process all observations simultaneously, as indicated by the formula (4.26), by observing that $O = P_k \wedge F_{k+1}$, we obtain the *two-filter* smoothing formula

$$\hat{o}^s_k = R \{ \hat{o}^f_{k|k} \wedge \hat{o}^b_{k|k+1} \} . \tag{4.27}$$

Interpretation: The above expression, through its use of estimation primitives and macros, provides a totally self-contained implementation of the two-filter smoother. However, for the sake of consistency with past work, it can be converted to the standard form of this smoother [18], [19] which expresses the smoothed estimate as a weighted combination of the forward and and backward Kalman filtering estimates. To do so, assume that x_k is estimable separately from P_k and F_{k+1}, and that the corresponding forward and backward error covariance matrices $\Sigma^f_{k|k}$ and $\Sigma^b_{k|k+1}$ are positive definite. The estimability assumption for x_k based on F_{k+1} is quite stringent since it requires that the dynamic matrix A_k should be invertible, which is often not true. Under these assumptions, the combined observation $\hat{o}^f_{k|k} \wedge \hat{o}^b_{k|k+1}$ can be expressed as

$$\begin{bmatrix} \hat{x}^f_{k|k} \\ \hat{x}^b_{k|k+1} \end{bmatrix} = \begin{bmatrix} I \\ I \end{bmatrix} x_k + \begin{bmatrix} \Sigma^{f\,1/2}_{k|k} & 0 \\ 0 & \Sigma^{b\,1/2}_{k|k+1} \end{bmatrix} \begin{bmatrix} u^f_k \\ u^b_k \end{bmatrix} . \tag{4.28}$$

The transformations which bring it to canonical form (where for convenience the nonzero blocks of SBV in (2.2) are not normalized to the identity) can be selected as

$$S_k = \begin{bmatrix} (\Sigma^f_{k|k})^{-1} & (\Sigma^b_{k|k+1})^{-1} \\ I & -I \end{bmatrix} \tag{4.29a}$$

$$V_k = \begin{bmatrix} (\Sigma^s_k)^{1/2} & 0 \\ 0 & (Q^s_k)^{-1/2} \end{bmatrix} \begin{bmatrix} (\Sigma^f_{k|k})^{-T/2} & (\Sigma^b_{k|k+1})^{-T/2} \\ \Sigma^{f\,1/2}_{k|k} & -\Sigma^{b\,1/2}_{k|k+1} \end{bmatrix} , \tag{4.29b}$$

with

$$(\Sigma_k^s)^{-1} \quad = \quad (\Sigma_{k|k}^f)^{-1} + (\Sigma_{k|k+1}^b)^{-1} \tag{4.30a}$$

$$Q_k^s \quad = \quad \Sigma_{k|k}^f + \Sigma_{k|k+1}^b . \tag{4.30b}$$

Note that V_k is orthonormal. This gives

$$\hat{o}_k^s \; : \quad (\Sigma_k^s)^{-1} \hat{x}_k^s = (\Sigma_k^s)^{-1} x_k + (\Sigma_k^s)^{-1/2} u_k^s \tag{4.31}$$

with

$$(\Sigma_k^s)^{-1} \hat{x}_k^s = (\Sigma_{k|k}^f)^{-1} \hat{x}_{k|k}^f + (\Sigma_{k|k+1}^b)^{-1} \hat{x}_{k|k+1}^b , \tag{4.32}$$

which constitute the usual two-filter smoothing formula.

C Double-sweep Smoother

To derive the double sweep or Rauch-Tung-Striebel (RTS) smoother [20], it is convenient to introduce the conditional observation

$$\omega_k(x_{k+1}) = X_{x_k, x_{k+1}} \{O\} \,|\, x_{k+1} \tag{4.33}$$

representing the information about x_k contained in all observations, conditioned on the knowledge of x_{k+1}. Taking into account the decomposition

$$O = P_k \wedge o_{k+1} \wedge F_{k+2} , \tag{4.34}$$

and noting that, given x_{k+1}, F_{k+2} does not contribute any information about x_k (x_k does not appear in this observation), we get

$$\omega_k(x_{k+1}) = (\hat{o}_{k|k}^f \wedge o_{k+1}) \,|\, x_{k+1} . \tag{4.35}$$

This expression represents the *forward sweep* of the RTS smoother. It can be implemented together with the forward Kalman recursion (4.7a).

To get the smoothed observation $\hat{o}_k^s$, we take into account the decomposition

$$R \circ X_{x_k, x_{k+1}} \{O\} = M_{x_{k+1}} \{\omega_k(x_{k+1})\} \wedge \hat{o}_{k+1}^s \tag{4.36}$$

with $\omega_k(x_{k+1}) \perp \hat{o}_{k+1}^s$, which in view of the definition (4.26) of a smoothed observation, implies

$$\hat{o}_k^s = R \circ X_{x_k} \{M_{x_{k+1}} \{\omega_k(x_{k+1})\} \wedge \hat{o}_{k+1}^s\} . \tag{4.37}$$

This recursion constitutes the *backward sweep* of the RTS smoother.

Interpretation: To gain a better understanding of the RTS recursions (4.35), (4.37), we derive below a semi-explicit form of the smoother which relies on finding closed-form implementations of the primitives appearing in the recursions. However closed-form realizations

of the primitives have the tendency to be less reliable numerically than the implicit imple-mentations presented earlier. Thus, the explicit form of the RTS smoother presented below is for illustrative purposes only.

(i) Forward sweep: Assume again that x_k is estimable from P_k. In addition, we assume that the noises $B_k u_k$ and $D_k u_k$ appearing in the dynamics d_{k+1} and measurements m_{k+1} are uncorrelated, i.e. $B_k D_k^T = 0$. In this case, given x_{k+1}, m_{k+1} does not contribute any information about x_k, so that

$$\omega_k(x_{k+1}) = (\hat{o}_{k|k}^f \wedge d_{k+1}) \,|\, x_{k+1} \,, \tag{4.38}$$

where $\hat{o}_{k|k}^f \wedge d_{k+1}$ is given by (4.12). Premultiplying (4.12) by

$$T_k = \begin{bmatrix} I & 0 \\ -A_k & I \end{bmatrix} \,, \tag{4.39}$$

and applying an orthonormal transformation V_k to the vector $[v_k^T \, u_k^T]^T$, where V_k is selected such that

$$\begin{bmatrix} \Sigma_{k|k}^{f\,1/2} & 0 \\ -A_k \Sigma_{k|k}^{f\,1/2} & B_k \end{bmatrix} V_k^T = \begin{bmatrix} G_k & H_k \\ 0 & J_k \end{bmatrix} \,, \tag{4.40}$$

the conditional observation $\omega_k(x_{k+1})$ can be expressed as

$$\omega_k(x_{k+1}) : \qquad \Theta_k \hat{x}_{k|k}^f + R_k x_{k+1} = x_k + G_k n_k \tag{4.41}$$

with

$$\Theta_k = I + H_k J_k^{-1} A_k \qquad R_k = -H_k J_k^{-1} E_{k+1} \,, \tag{4.42}$$

where $n_k \sim \mathcal{N}(0, I)$. To interpret the matrices Θ_k, R_k, and G_k appearing in (4.41), note that by multiplying (4.40) by its transpose, we find

$$J_k J_k^T = A_k \Sigma_{k|k}^f A_k^T + B_k B_k^T \triangleq \Sigma_{k+1|k}^f \tag{4.43a}$$

$$H_k J_k^T = -\Sigma_{k|k}^f A_k^T \tag{4.43b}$$

$$G_k G_k^T = \Sigma_{k|k}^f - \Sigma_{k|k}^f A_k^T (\Sigma_{k+1|k}^f)^{-1} A_k \Sigma_{k|k}^f \triangleq \Sigma_k^p \,. \tag{4.43c}$$

In the above expressions, we have assumed that the matrix J_k is invertible. But, according to (4.43a), J_k is a square-root of the covariance matrix $\Sigma_{k+1|k}^f$ of the noise $B_k u_k - A_k \Sigma_{k|k}^{f\,1/2} v_k$ in the predicted observation $\hat{o}_{k+1|k}^f$. Thus, J_k will be invertible whenever $\Sigma_{k+1|k}^f$ is positive definite. Substituting (4.43a) inside the expressions (4.42) for Θ_k and R_k gives

$$\Theta_k = I - \Sigma_{k|k}^f A_k^T (\Sigma_{k+1|k}^f)^{-1} A_k \tag{4.44a}$$

$$R_k = \Sigma_{k|k}^f A_k^T (\Sigma_{k+1|k}^f)^{-1} E_{k+1} \,. \tag{4.44b}$$

Finally, from (4.41), we see that $G_k G_k^T = \Sigma_k^p$ is the filtered error variance for x_k based on the observation $\omega_k(x_{k+1})$, i.e based on the past observations, and the "pinned" value of x_{k+1}.

Backward sweep: Since the mutation of x_{k+1} inside $\omega_k(x_{k+1})$ is accomplished by moving the $R_k x_{k+1}$ term to the right hand side, the observation $M_{x_{k+1}}\{\omega_k(x_{k+1})\} \wedge \hat{o}_{k+1}^s$ takes the form

$$\begin{bmatrix} \Theta_k \hat{x}_{k|k}^f \\ \hat{x}_k^s \end{bmatrix} = \begin{bmatrix} I & -R_k \\ 0 & I \end{bmatrix} \begin{bmatrix} x_k \\ x_{k+1} \end{bmatrix} + \begin{bmatrix} G_k & 0 \\ 0 & \Sigma_{k+1}^{s\,1/2} \end{bmatrix} \begin{bmatrix} n_k \\ u_{k+1}^s \end{bmatrix} . \qquad (4.45)$$

To extract x_k from this observation, we need only to premultiply it by $[\,I\ R_k\,]$, which gives (4.31) with

$$\hat{x}_k^s \quad = \quad \Theta_k \hat{x}_{k|k}^f + R_k \hat{x}_{k+1}^s \qquad (4.46a)$$

$$\Sigma_k^s \quad = \quad R_k \Sigma_{k+1}^s R_k^T + \Sigma_k^p . \qquad (4.46b)$$

D   Application to Failure Detection

The descriptor Kalman filtering recursions of Section 4.A are directly applicable to a recently proposed [21] failure detection procedure for dynamical systems containing nuisance parameters, i.e., quantities taking arbitary unknown values. Given a standard state-space model where a nuisance term appears additively in the dynamics, the nuisance can be removed by premultiplying the dynamics with a matrix which nulls out the nuisance space. This yields a system with descriptor dynamics, where the presence or absence of failures can be determined by applying either innovations–based statistical failure detection tests, or algebraic tests relying on parity checks. Since the Kalman filtering recursions (4.7a)–(4.7c) generate both innovations and parity checks, they provide all the data necessary to implement failure detection schemes.

Our objective at this stage is not to discuss failure detection issues in detail, but just to illustrate the applicability of descriptor Kalman filtering algorithms. Accordingly, we follow the formulation of reference [21], to which the reader is referred for more extensive discussions and a comprehensive bibliography.

As starting point, we formulate the standard binary hypothesis testing problem for linear observation models in terms of the primitives introduced earlier. Given the measurement vector y, and two models

$$\mathcal{H}_0 \quad : \quad Ey = Bu \qquad (4.47b)$$

$$\mathcal{H}_1 \quad : \quad Ey = Ax + Bu , \qquad (4.47b)$$

we seek to decide which one best fits y. Such a decision requires

(i) checking whether the constraints $\mathcal{S}\{\mathcal{H}_0\}$ and/or $\mathcal{S}\{\mathcal{H}_1\}$ are satisfied, and, if both are satisfied, then

(ii) computing the log-likelihood ratio $\log L\{\mathcal{H}_1\} - \log L\{\mathcal{H}_0\}$ for comparison to a selected threshold.

Performing this decision reduces therefore to computing

$$\overline{R}\{\mathcal{H}_0\} \quad \text{and} \quad \overline{R}\{\mathcal{H}_1\}. \tag{4.48}$$

A careful inspection of the canonical decomposition algorithm following Lemma 2.1 reveals that the numerical procedure required to compute $\overline{R}\{\mathcal{H}_1\}$ yields $\overline{R}\{\mathcal{H}_0\}$ *at the same time.*

Next, suppose we consider the same problem as before, but with an unknown nuisance vector n, so that we need to decide between $\mathcal{H}_0$ and $\mathcal{H}_1$, where

$$\mathcal{H}_0 \quad : \quad Ey = Gn + Bu \tag{4.49a}$$

$$\mathcal{H}_1 \quad : \quad Ey = Ax + Gn + Bu. \tag{4.49b}$$

In this case, all we need to do is first to extract x from the observations $\mathcal{H}_i$ with $i = 0, 1$, which has the effect of removing the nuisance term Gn, and then apply the previous hypothesis testing procedure to the extracted observations $X_x\{\mathcal{H}_i\}$. Again, because X_x requires only the computation of the left null space of G, a single algorithm can be used to generate $X_x\{\mathcal{H}_i\}$ for $i = 0, 1$.

Finally, let us examine the failure detection problem with isolation studied in [21]. We are given the system

$$x_{k+1} \quad = \quad Ax_k + Bv_k + G_x n_k + \Gamma u_k + f m_k \tag{4.50a}$$

$$y_k \quad = \quad Cx_k + Dv_k + G_y n_k, \tag{4.50b}$$

where x_k is the state of the system, u_k the known input, y_k the observed output, and v_k a white Gaussian noise sequence such that $v_k \sim \mathcal{N}(0, I)$. The vector

$$\begin{bmatrix} G_x \\ G_y \end{bmatrix} n_k$$

represents a *nuisance* modelling unobserved input and output disturbances or possible alternative failures not considered here. The term $f m_k$ represents the actuator failure that we seek to detect and isolate. For simplicity, it is assumed that the matrices apearing in the above dynamics are *constant*, and have appropriate dimensions. In the above model, the absence of failure corresponds to the case where $f = 0$, whereas in the presence of a failure, the known vector $f \neq 0$. To cast the above model within our framework, we rewrite (4.50a)–(4.50b) as the observation

$$\mathcal{H}_k^i \quad : \quad \begin{bmatrix} -\Gamma u_k \\ y_k \end{bmatrix} = \begin{bmatrix} -I & A & G_x & f \\ 0 & C & G_y & 0 \end{bmatrix} \begin{bmatrix} x_{k+1} \\ x_k \\ n_k \\ m_k \end{bmatrix} + \begin{bmatrix} B \\ D \end{bmatrix} v_k. \tag{4.51}$$

The indices k and i appearing in $\mathcal{H}_k^i$ correspond respectively to the time index and hypothesis. Thus, $i = 0$ in the absence of failure, and $i = 1$ when a failure is present. The onset of a failure corresponds to i jumping from 0 to 1 at a certain time s. We seek to detect such a jump, and estimate the failure time s. This can be accomplished by applying the following algorithm.

1) For $i = 0, 1$ and $k \geq 0$, compute

$$X_{(x_{k+1}, x_k, m_k)}\{\mathcal{H}_k^i\}\,.$$

Because the matrices G_x and G_y do not depend on either the choice of hypothesis or the time index k, the above extraction needs to be performed *only once* for $i = 0$, 1, and all $k \geq 0$; see the previous comment concerning the binary hypothesis problem with nuisance rejection.

2) Apply the Kalman filtering recursions (4.7a)–(4.7c), (4.8) to the observation stream

$$O(s, k) = \left(\bigwedge_{l=0}^{s} X_{x_{l+1}, x_l, m_l}\{\mathcal{H}_l^0\} \right) \wedge \left(\bigwedge_{l=s+1}^{k} X_{x_{l+1}, x_l, m_l}\{\mathcal{H}_l^1\} \right) \qquad (4.52)$$

which models a situation where a failure has occured at time $s \leq k$ (the case $s = k$ corresponds to the absence of failure). Then, verify whether the measurement vectors belong to the constraint space $\mathcal{S}\{O(s, k)\}$, and compute

$$\frac{\max_{k-L < s \leq k} L(O(s, k))}{L(O(k, k))} \quad \text{and} \quad \hat{s} = \arg \max_{k-L < s \leq k} L(O(s, k)) \qquad (4.53)$$

for a properly chosen window length L. When the likelihood ratio exceeds a properly selected threshold, a failure occuring at time $\hat{s}$ has been detected.

Note that if the system under consideration is an interconnection of local subsystems, and the nuisance affects only a subsystem, the nuisance rejection of step 1) can be performed locally. However, the Kalman filtering phase of the algorithm remains global.

5   Finite State Processes

The previous results can be extended to the nonlinear nonGaussian case, when the parameter vector X to be estimated belongs to a finite set $\mathcal{X}$.

A   Observation Model

An observation o consists as before of a pair $(\mathcal{C}, \mathbf{P})$ formed by a set $\mathcal{C}$ defining the allowable pairs of measurements Y and parameter vector X, and, depending on whether the measurement vector Y takes discrete or continuous values, a probability distribution

$$\mathbf{P}(Y = y, X = x) = \mathbf{p}(y, x)\,, \qquad (5.1)$$

or a density

$$\mathbf{P}(Y \in d\mathcal{S}, X = x) = \mathbf{p}(y, x)dV \tag{5.2}$$

with $(y, x) \in \mathcal{C}$, where $dV = |d\mathcal{S}|$ represents an element of volume of the set $\mathcal{S}$ of measurements Y compatible with the constraint set $\mathcal{C}$. If $\mathcal{Y}$ and $\Pi_{\mathcal{Y}}$ denote respectively the domain of y, and the projection operator which associates to a pair (y, x) its coordinate y, i.e.

$$\Pi_{\mathcal{Y}}(y, x) = y , \tag{5.3}$$

the set $\mathcal{S}$ of valid outputs is given by

$$\mathcal{S} = \Pi_{\mathcal{Y}} \mathcal{C} . \tag{5.4}$$

The reference measure dV which is defined on $\mathcal{S}$ is left unspecified, since it depends on the set $\mathcal{S}$. Finally, given $y \in \mathcal{S}$, the set $\mathcal{X}(y)$ represents the set of parameter vectors x such that the pair $(y, x) \in \mathcal{C}$.

Since X belongs to a finite set, we shall focus our interest exclusively on ML estimates. As before, we say that X is estimable if $\mathbf{p}(y, x)$ admits a unique maximum over $\mathcal{X}(y)$, in which case the ML estimate is defined as

$$\hat{x}_{ML}(y) = \arg \max_{x \in \mathcal{X}(y)} \mathbf{p}(y, x) . \tag{5.5}$$

The corresponding likelihood and conditional likelihood functions are then given by

$$\mathbf{q}(y) = \mathbf{p}(y, \hat{x}_{ML}(y)) = \max_{x \in \mathcal{X}(y)} \mathbf{p}(y, x) \tag{5.6}$$

$$\mathbf{q}(x|y) = \frac{\mathbf{p}(y, x)}{\mathbf{q}(y)} , \tag{5.7}$$

which are defined respectively over $\mathcal{S}$ and $\mathcal{X}(y)$. Thus, as in the Gaussian case, we have decomposed the pair $(\mathcal{C}, \mathbf{p}(y, x))$ into two subpairs $(\mathcal{S}, \mathbf{q}(y))$ and $(\mathcal{X}(y), \mathbf{q}(x|y))$ describing the marginal and conditional constraint sets and likelihood functions. Again, $\mathbf{q}(y)$ will be denoted as $L(o)$ to indicate its association with o.

However, unlike the Gaussian case, the likelihood and conditional likelihood functions $\mathbf{q}(y)$ and $\mathbf{q}(x|y)$ do not coincide with the marginal and conditional densities

$$\mathbf{p}(y) = \sum_{x \in \mathcal{X}(y)} \mathbf{p}(y, x) \tag{5.8}$$

$$\mathbf{p}(x|y) = \frac{\mathbf{p}(y, x)}{\mathbf{p}(y)} \tag{5.9}$$

which differ from (5.6) and (5.7) by the fact that we are summing instead of maximizing over x.

B Primitives

The primitives introduced in Section 3 for linear Gaussian observations can be generalized as follows. First, consider two observations $o_i = (\mathcal{C}_i, \mathbf{P}_i)$ with $i = 1, 2$, where the parameter vectors of o_1 and o_2 can be decomposed into attributes x_1 and x_2 which are specific to these observations, as well as a component x_c common to both. Similarly the measurements vectors of o_1 and o_2 contain parts y_1 and y_2 which are specific to these observations, as well as a common part y_c. Then, adopting the notation (3.3a)–(3.3c) for the measurement and parameter vectors of each observation and their combination, the combined observation $o_1 \wedge o_2$ is formed by the pair $(\mathcal{C}, \mathbf{P})$ such that

$$
\begin{aligned}
\mathcal{C} &= \mathcal{C}_1 \wedge \mathcal{C}_2 \\
&= \{\, (y, x) \; : \; (\eta_1, \chi_1) \in \mathcal{C}_1 \text{ and } (\eta_2, \chi_2) \in \mathcal{C}_2 \,\} \tag{5.10} \\
\mathbf{p}(y, x) &= \mathbf{p}_1(\eta_1, \chi_1)\mathbf{p}_2(\eta_2, \chi_2) \,. \tag{5.11}
\end{aligned}
$$

Because the probability distributions/densities we are considering are nonGaussian with finite valued parameter vectors, both $\mathbf{p}(y, x)$ and $\mathbf{q}(x|y)$ are minimal sufficient statistics for estimating X, i.e. we cannot base the reduction operation $R\{.\}$ on the need to extract a minimal sufficient observation from o. However, as indicated by (3.14), an important property of the reduction operation $R\{.\}$ in the Gaussian case is that it retains no information about the output y, in the sense that its likelihood function

$$
L(R\{o\}) = 1 \tag{5.12}
$$

is uniformly distributed. The conditional likelihood function $\mathbf{q}(x|y)$ has precisely this characteristic, i.e.

$$
\max_x \mathbf{q}(x|y) = 1 \,. \tag{5.13}
$$

This implies that the primitives $R\{.\}$ and $\overline{R}\{.\}$ need to be defined as:

$$
\begin{aligned}
R\{o\} &= (\mathcal{C}_R, \mathbf{q}(x|y)) \tag{5.14} \\
\overline{R}\{o\} &= (\mathcal{S}, \mathbf{q}(y)) \,, \tag{5.15}
\end{aligned}
$$

where the set $\mathcal{C}_R$ is obtained by removing from $\mathcal{C}$ the constraints affecting y alone, so that

$$
\mathcal{C} = \mathcal{C}_R \wedge \mathcal{S} \,. \tag{5.16}
$$

Note that because the conditional likelihood function $\mathbf{q}(x|y)$ given by (5.7) is defined only on $\mathcal{C}$, the definition of $R\{o\}$ in (5.14) requires extending $\mathbf{q}(x|y)$ from $\mathcal{C}$ to the larger set $\mathcal{C}_R$, where the choice of extension does not matter, since only measurement vectors y in $\mathcal{S}$ (and thus pairs (x, y) in $\mathcal{C}$) will ultimately be allowed.

The mutation corresponds again to the change of status of a parameter or measurement from unknown to known, or vice versa, and the conditional observation and extraction macros are defined by (3.28a)–(3.28b), so that for an observation o with density $\mathbf{p}(y, x_0, x_1)$, the density of the extracted observation $X_{x_1}\{o\}$ is given by

$$
\mathbf{q}(y, x_1) = \max_{x_0} \mathbf{p}(y, x_0, x_1) \,. \tag{5.17}
$$

C   Estimation of Hidden Markov Models

To illustrate the application of the above macros, we consider the problem of finding the most likely state sequence of a hidden Markov model (HMM). In such models, the states of a Markov chain are not observed directly, but need to be estimated from a sequence of measurements depending indirectly on the states. Models of this type are often used in speech processing [8], or in digital communications for decoding convolutional codes or the deconvolution of intersymbol interference [22], [23], [24].

Let x_k be a Markov chain defined for $0 \le k \le N$, taking values in a finite set $\mathcal{X}$. The joint probability distribution of the chain can be expressed as

$$\mathbf{p}(X) = \mathbf{q}_i(x_0)\mathbf{q}_e(x_N) \prod_{k=0}^{N-1} \tau(x_k, x_{k+1}) \,. \tag{5.18}$$

This form is a slight variant of the standard expression for Markov processes, which is sometimes employed [25] to model Markov processes for which some information about both the initial and final state is available. This situation occurs for example when we use a convolutional encoder to transmit data blocks separated by zero bits whose role is to ensure that the encoder starts and terminates in a known state (the zero state). For such a case, we have

$$\mathbf{q}_i(x) = \mathbf{q}_e(x) = \delta(x) \,, \tag{5.19}$$

where $\delta(x)$ denotes the distribution equal to one for $x = 0$, and zero otherwise.

The observations y_k with $0 \le k \le N - 1$ may take either continuous or discrete values. Given a complete state trajectory, they are modeled by the conditional density/distribution

$$\mathbf{p}(Y|X) = \prod_{k=0}^{N-1} \phi(y_k, x_k, x_{k+1}) \,, \tag{5.20}$$

whose form indicates that y_k may be either a measurement of the state x_k or of the transition from x_k to x_{k+1}. For simplicity, we assume that there are no constraints on the states x_k and observations y_k, so that an observation o will consist only of a probability distribution/density.

Then the joint distribution/density for the measurements and states can be expressed as

$$\mathbf{p}(Y, X) = \mathbf{p}(Y|X)\mathbf{p}(X) = \bigwedge_{j=0}^{N+1} o_j \,, \tag{5.21}$$

where

$$o_{k+1} \quad : \quad \tau(x_k, x_{k+1})\phi(y_k, x_k, x_{k+1}) \tag{5.22}$$

for $0 \le k \le N - 1$, and

$$o_0 \quad : \quad \mathbf{q}_i(x_0) \qquad o_{N+1} \quad : \quad \mathbf{q}_e(x_N) \,. \tag{5.23}$$
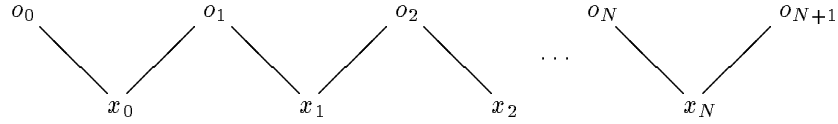
Figure 1: Linear XO-graph.

The maximum likelihood sequence estimation problem for HMMs consists in jointly maximizing $\mathbf{p}(Y, X)$ with respect to the variables x_k with $0 \le k \le N$. The best known solution of this problem is the celebrated Viterbi algorithm [26], [22]. To relate this algorithm to those discussed in Section 4 for descriptor systems, it is useful to construct the XO-graph associated to the observations (5.22)–(5.23). The concept of XO-graph was introduced in [7] for linear Gaussian observations, but it applies also to the nonGaussian case. Its role is to display the dependency structure of the observations with respect to the variables to be estimated. Conceptually, it is similar to the graph representing the neighborhood structure of a Markov random field [9]. In such a graph, the observations o_j and variables x_k are represented respectively by o- and x- nodes. An edge connects the nodes o_j and x_k if the variable x_k appears in the probability distribution/density corresponding to o_j. For the HMM considered here, o_{k+1} is connected only to the nodes x_k and x_{k+1}, so that the HMM (5.22)–(5.23) admits the linear graph structure shown in Fig. 1.

But the descriptor estimation problem of Section 4 admits exactly the same XO-graph! Since the structure of this graph determines completely the order in which observations are processed and the variables x_k are estimated, we can immediately conclude that *the filtering and smoothing algorithms described in Section 4 are directly applicable to the HMM considered here.* However, since the primitives are not implemented in the same manner, it is of interest to examine more precisely the internal mechanics and outcome of each algorithm.

Forward pass: Let P_k be the combination (4.5) of all observations in the past of time k. Then, the filtered observation $\hat{o}^f_{k|k}$ given by (4.6a) corresponds actually to the function $\mathbf{q}^f_k(x_k)$ measuring the likelihood of the best path terminating in state x_k given the measurements $y_1, \ldots, y_{k-1}$. Because the definition of $\hat{o}^f_{k|k}$ includes a reduction, the distribution $\mathbf{q}^f_k(x_k)$ is normalized, i.e. its maximum value over all x_k's is one. A similar normalization is often performed in convolutional decoders by subtracting periodically a constant from the value function $J^f_k(x_k) = -\ln \mathbf{q}^f_k(x_k)$ to prevent it from growing excessively. This is possible because we are only interested in the relative value of J^f_k or $\mathbf{q}^f_k$ for different x_k's. Then, if we consider the recursion (4.7a), the extracted observation $X_{x_{k+1}}\{\hat{o}^f_{k|k} \wedge o_{k+1}\}$ is specified by the distribution

$$\xi^f_{k+1}(x_{k+1}) = \max_{x_k} \mathbf{q}^f_k(x_k)\tau(x_k, x_{k+1})\phi(y_k, x_k, x_{k+1}), \qquad (5.24a)$$

where the argument of the maximum is given by

$$x_k^* = \psi_{k+1}(x_{k+1}), \tag{5.24b}$$

and the dependence of x_k^* on the known observation y_k has been suppressed. Given that the state of the Markov chain at time $k+1$ is x_{k+1} and the observations up to time k are y_0, ... y_k, $x_k^* = \psi_{k+1}(x_{k+1})$ represents the most likely state at time k. The function $\psi_{k+1}(.)$ is usually called the back-pointer function, since given the current state at time $k+1$, it points to the most likely state at the previous instant.

Then, if

$$\begin{aligned} \lambda_{k+1} &= \overline{R}_{\circ} X_{x_{k+1}} \{ \hat{o}_{k|k}^f \wedge o_{k+1} \} \\ &= \max_{x_{k+1}} \xi_{k+1}^f(x_{k+1}), \end{aligned} \tag{5.25}$$

we have

$$\mathbf{q}_{k+1}^f(x_{k+1}) = \xi_{k+1}^f(x_{k+1})/\lambda_{k+1}. \tag{5.26}$$

Furthermore, according to (4.7c), if L_k denotes the likelihood function of the observation P_k, it satisfies the recursion

$$L_{k+1} = \lambda_{k+1} L_k, \tag{5.27}$$

which can be used to evaluate the likelihood function of the measurement sequence.

According to (4.8), the forward filter can be initialized with

$$L_0 = \lambda_0 = \max_{x_0} \mathbf{q}_i(x_0) \tag{5.28a}$$

$$\mathbf{q}_0^f(x_0) = \mathbf{q}_i(x_0)/\lambda_0. \tag{5.28b}$$

Viterbi/double sweep decoder: If O represents the combined observation (5.21), the smoothed observation $\hat{o}_k^s$ given by (4.26) is the distribution $\mathbf{q}_k^s(x_k)$ obtained after maximizing $\mathbf{p}(Y, X)$ with respect to all x_l with $l \neq k$. Its maximum

$$\hat{x}_k = \arg\max_{x_k} \mathbf{q}_k^s(x_k) \tag{5.29}$$

represents the most likely state at time k, i.e. the most likely trajectory corresponding to the measurements y_k with $0 \leq k \leq N-1$ goes through $\hat{x}_k$ at time k.

From the definition of the conditional observation macro, we find that the observation $\omega_k(x_{k+1}) = \hat{o}_{k|k}^f \wedge o_{k+1} \mid x_{k+1}$ admits the distribution

$$\mathbf{q}_k^f(x_k)\tau(x_k, x_{k+1})\phi(y_k, x_k, x_{k+1})/\mathbf{q}_{k+1}^f(x_{k+1}) \tag{5.30}$$

so that according to (4.36), the observation $R_{\circ} X_{x_k, x_{k+1}} \{O\}$ has for distribution

$$\mathbf{q}_k^f(x_k)\tau(x_k, x_{k+1})\phi(y_k, x_k, x_{k+1})\mathbf{q}_{k+1}^s(x_{k+1})/\mathbf{q}_{k+1}^f(x_{k+1}). \tag{5.31}$$

Its joint maximum over x_k and x_{k+1} gives the most likely states $\hat{x}_k$ and $\hat{x}_{k+1}$ at times k and $k+1$. Maximizing over x_k first, and noting that the x_k dependent part of (5.31) is identical to the one minimized in (5.24a) yields

$$\hat{x}_k = \psi_{k+1}(\hat{x}_{k+1}) \ . \tag{5.32}$$

Thus, the most likely trajectory is obtained through a double sweep algorithm, where the forward sweep computes the trace back function $\psi_{k+1}(.)$, and the backward sweep generates the optimum state trajectory, starting from the most likely final state $\hat{x}_N$. This state can be computed by observing that the observation $\hat{o}_N^s = \hat{o}_{N|N}^f \wedge o_{N+1}$ has for distribution

$$\mathbf{q}_N^s(x_N) = \mathbf{q}_N^f(x_N)\mathbf{q}_e(x_N) \ , \tag{5.33}$$

whose maximum yields $\hat{x}_N$.

By employing primitives which are sensitive to data type, i.e. which act differently on Gaussian and finite state processes, we have therefore been able to show that the Viterbi and Rauch-Tung-Striebel double-sweep algorithms are almost identical, as was first noted informally in [10]. Of course, one could push forward the analogy existing between the ML estimation problems for descriptor systems and HMMs by attempting to develop a "two-filter" decoder for the most-likely state sequence of a HMM. However, such a decoder would be less attractive than the Viterbi decoder, since it would require the solution of two dynamic programming problems, one for each time direction.

6 Conclusions

In this paper we have developed a set of high level primitives for ML estimation which, because they rely on general statistical concepts, are applicable to estimation problems where sufficient statistics admit finite parametrizations, such as linear Gaussian models or finite-state processes. The power of these high level primitives has been illustrated by considering the ML estimation problems for linear descriptor systems and hidden Markov models.

The work presented here can be extended in at least two directions. First, for the finite-state case, we have assumed that observations can be modeled by a distribution/density and constraint set pair $(\mathbf{p}(y, x), \mathcal{C})$. However for certain applications, such as queing networks, the sigma field of events that can be probabilized is more restricted, so that an observation needs to be modeled as a triple $(\Omega, \mathcal{A}, \mathbf{P})$ formed by a sample space, sigma field, and probability measure. In this case, because of the possibility that different observations might have common components, certain technical difficulties arise in the definition of the combinator $\wedge$, which are examined in [6].

Another area where the primitives developed in this paper are expected to be useful is for the formulation of stochastic dynamic programming problems involving either linear Gaussian systems or controlled Markov chains.

A third direction of research concerns the large amount of flexibility existing in the choice of recursive strategies for solving ML estimation problems specified by coupled observations. For the smoothing problem of descriptor systems, the two-filter and double-sweep smoothers are smoothers are only two of several possible processing strategies. A third procedure consists for example of processing the observations outward from the center of the interval, until the edges are reached, and then performing a second inward pass. The choice of possible algorithms becomes even larger when a network of interconnected descriptor systems is considered, since in this case locality can be exploited both in time and in space. This raises therefore the question of finding a way to describe all processing strategies within a single framework, in order to be able to select optimum algorithms in terms of memory requirements and computational cost.

Acknowledgements

References

[1]   H. W. Sorenson, *Kalman Filtering: Theory and Application.* New York: IEEE Press, 1985.

[2]   B. D. O. Anderson and J. B. Moore, *Optimal Filtering.* Englewood Cliffs, NJ: Prentice–Hall, 1979.

[3]   G. J. Bierman, *Factorization Methods for Discrete Sequential Estimation.* New York: Academic Press, 1977.

[4]   T. Kailath, *Lectures on Wiener and Kalman Filtering.* Berlin: Springer Verlag, 1985.

[5]   P. Le Guernic, T. Gauthier, M. Le Borgne, and C. Le Maire, "Programming real-time applications with Signal" *Proc. IEEE*, vol. 79, pp. 1321–1336, Sep. 1991.

[6]   A. Benveniste, "Constructive probability and the Signalea language: building and handling random processes with programming," Tech. Rep. 1532, Institut National de Recherche en Informatique et Automatique, Rocquencourt, France, Oct. 1991.

[7]   R. Nikoukhah, D. Taylor, B. C. Levy, and A. S. Willsky, "Graph structure and recursive estimation of noisy linear relations," Tech. Rep. 1912, Institut National de Recherche en Informatique et Automatique, Rocquencourt, France, May 1993.

[8]   L. R. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, pp. 4–16, Jan. 1986.

[9] R. Kindermann and J. L. Snell, *Markov Random Fields and their Applications*. Providence, RI: American Mathematical Society, 1980.

[10] B. Delyon, "Remarks on filtering of semi-Markov data," Tech. Rep. 733, Institut National de Recherche en Informatique et Systèmes Aléatoires, Rennes, France, May 1993.

[11] E. L. Lehmann, *Theory of Point Estimation*. Pacific Grove, CA: Wadsworth & Brooks Cole, 1991.

[12] X. Wang and P. Bernhard, "Filtrage et lissage des systèmes discrets," Tech. Rep. 1083, Institut National de Recherche en Informatique et Automatique, Rocquencourt, France, Aug. 1989.

[13] R. Nikoukhah, A. S. Willsky, and B. C. Levy, "Kalman filtering and Riccati equations for descriptor sytems," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1325–1342, Sep. 1992.

[14] L. Chisci and G. Zappa, "Square-root Kalman filtering of descriptor systems," *Systems Control Lett.*, vol. 19, pp. 325–334, Oct. 1992.

[15] D. Taylor and A. S. Willsky, "Maximum likelihood estimation for two-point boundary-value descriptor systems," in *Proc. 1991 Conf. Information Sciences and Systems*, March 1991.

[16] A. S. Willsky, "A survey of design methods for failure detection in dynamic systems," *Automatica*, vol. 12, pp. 601–611, 1976.

[17] E. Y. Chow and A. S. Willsky, "Analytical redundancy and the design of robust failure detection systems," *IEEE Trans. Automat. Contr.*, vol. 29, pp. 603–614, July 1984.

[18] D. Q. Mayne, "A solution of the smoothing problem for linear dynamic systems," *Automatica*, vol. 4, pp. 73–92, Nov. 1966.

[19] D. C. Fraser, *A New Technique for the Optimal Smoothing of Data*. PhD thesis, Dept. Aeronautics and Astronautics, M.I.T., Cambridge, MA, Jan. 1967.

[20] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear systems," *AIAA Journal*, vol. 3, pp. 1445–1450, Aug. 1965.

[21] A. Benveniste, M. Basseville, R. Nikoukhah, and A. S. Willsky, "On the use of descriptor systems for failure detection and isolation," in *Proc. IFAC World Congress*, (Sidney, Australia), pp. 497–500, July 1993.

[22] G. D. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268–278, March 1973.

[23] G. D. Forney, "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interferences," *IEEE Trans. Inform. Theory*, vol. 18, pp. 363–378, May 1972.

[24] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding.* New York: McGraw-Hill, 1979.

[25] B. Jamison, "Reciprocal processes," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 30, pp. 65–86, 1974.

[26] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotic optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 13, pp. 260–269, Apr. 1967.

INRIA