Graph structure and recursive estimation of noisy linear relations

Ramine Nikoukhah, D. Taylor, Bernard C. Levy, A.S. Willsky

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Graph structure and recursive estimation of noisy linear relations

Ramine NIKOUKHAH - Darrin TAYLOR
Bernard C. LEVY - Alan S. WILLSKY

N° 1912
Mai 1993

PROGRAMME 5

Traitement du signal,
automatique et
productique

GRAPH STRUCTURE AND RECURSIVE ESTIMATION OF NOISY LINEAR RELATIONS

Ramine Nikoukhah, *INRIA, Rocquencourt*
Darrin Taylor, MIT, Cambridge, MA, USA
Bernard C. Levy, UCD, CA, USA
Alan S. Willsky, MIT, Cambridge, MA, USA

This paper examines estimation problems specified by noisy linear relations describing either dynamical models or measurements. Each such problem has a graph structure, which can be exploited to derive recursive estimation algorithms only when the graph is acyclic, i.e., when it is obtained by combining disjoint trees. Aggregation techniques appropriate for reducing an arbitrary graph to an acyclic one are presented. The recursive maximum likelihood estimation procedures that we present are based on two elementary operations, called reduction and extraction, which are used to compress successive observations, and discard unneeded variables. These elementary operations are used to derive filtering and smoothing formulas applicable to both linear and arbitrary trees, which are, in turn applicable to estimation problems in settings ranging from 1-D descriptor systems to 2-D difference equations to multiscale statistical models of random fields. These algorithms can be viewed as direct generalizations to a far richer setting of Kalman filtering and both two-filter and Rauch-Tung-Striebel smoothing for standard causal state space models.

STRUCTURE DE GRAPHE ET ESTIMATION RÉCURSIVE DE RELATIONS LINÉAIRES BRUITÉES

Ce rapport étudie les problèmes d'estimation spécifiés par des relations linéaires bruitées décrivant des modèles dynamiques ou des mesures. A chaque problème de ce type est associée une structure de graphe, laquelle ne peut être exploitée pour générer des algorithmes récursifs que lorsque le graphe est acyclique, i.e. quand il correspond à la juxtaposition d' arbres disjoints. Des méthodes d'agrégation permettant de transformer un graphe arbitraire en un graphe acyclique sont presentées. La méthode d' estimation récursive par maximum de vraisemblance que nous décrivons se fonde sur deux opérations élémentaires, la réduction et l'extraction, qui sont employées pour comprimer les observations successives, et éliminer les variables sans interêt. Ces opérations élémentaires fournissent des formules de filtrage et de lissage applicables aussi bien aux arbres linéaires qu'aux arbres généraux. De tels graphes apparaissent dans des problèmes d' estimation allant des systèmes descripteurs à une dimension aux équations aux différences à deux dimensions, en passant par des modèles multiéchelles de champs aléatoires. Ces algorithmes peuvent être considérés comme la généralisation directe dans un cadre beaucoup plus large du filtrage de Kalman ansi que des formules des deux filtres ou de Rauch-Tung-Striebel pour le lissage des modèles par espace d'états.

1 Introduction

In this paper we investigate the problem of recursive estimation for a set of unknown variables subject to noisy, linear, constraints. Our motivation for this is to provide a unifying framework that includes not only the standard, causal Kalman filter and its information filter counterpart but also applies equally well to a much richer set of problems in which some of the natural, simplifying aspects of the standard problem, that are usually taken for granted, don't apply, requiring more careful analysis.. To understand part of our perspective, consider for the moment the standard Kalman filtering problem

$$x_{k+1} = A_k x_k + B_k w_k \quad , \quad k \geq 0 \tag{1}$$

$$y_k = C_k x_k + \tau_k \quad , \quad k \geq 1 \tag{2}$$

where w_k and v_k are independent, zero-mean Gaussian random vectors with identity covariances and where x_0 is a Gaussian random vector, independent of w and v, with mean m_0 and covariance P_0. While there are a variety of ways in which to derive optimal estimation algorithms for this standard problem, as we will see, the one that we must use in our general case involves adopting a maximum likelihood (ML) perspective in which initial conditions, dynamics (1) and observation (2) are all viewed as "measurements", or perhaps, more appropriately, as noisy dynamic constraints. That is, we wish to estimate the sequence of unknowns $x_0, x_1, \cdots$ from the sequence of measurements

$$\begin{bmatrix} m_0 \\ 0 \\ y_1 \\ 0 \\ y_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} I & 0 & 0 & \cdots \\ -A_0 & I & 0 & \cdots \\ 0 & C_1 & 0 & \cdots \\ 0 & -A_1 & I & \cdots \\ 0 & 0 & C_2 & \\ \vdots & \vdots & \vdots & \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ \vdots \\ \vdots \end{bmatrix} + \begin{bmatrix} \tilde{x}_0 \\ B_0 w_0 \\ \tau_1 \\ B_1 w_1 \\ \tau_2 \\ \vdots \end{bmatrix} \tag{3}$$

where $\tilde{x}_0 = x_0 - m_0$.

Note three things about this formulation First of all, thanks to the lower-bi-diagonal structure of the matrix on the right-hand side of (3), we can readily, obtain recursive algorithms for the computation of the desired filtered or smoothed estimates, essentially by Gaussian elimination. Secondly, in many cases the dynamic noise $B_k w_k$ is *not* full rank (e.g. think about a second-order system driven by a single noisy input). Consequently if we adopt the ML persective we apparently have to deal with a singular estimation problem, since the "measurement noise" in (3) has a singular covariance. Thirdly, thanks to the identity blocks in the matrix in (3), it is easy to see that the matrix relating $x_0, \cdots, x_k$ to the measurements in (3) has full column rank implying that this ML estimation problem is well-posed in that we do indeed have information about the full state vector. However, in some situations it is not realistic to assume that we have useful prior information about x_0 so that we either eliminate the first block row of (3) or replace it by only partial prior information about x_0. In this case, of course, the matrix in (3) may no longer have full column rank.

In this standard framework the two apparent sources of singularity that we have pointed out typically cause us no real difficulty. In particular, thanks again to the presence of the identity blocks in (3) – or more fundamentally to the recursive nature of the dynamic constraint in (1), the calculations corresponding to the incorporation of these dynamics in the estimation procedure are reduced to the essentially trivial prediction step of the Kalman filter, so that the singularity of $B_k B_k^T$ causes no difficulty. Also, if x_0 is partially or completely unknown, we can use the information form of the filter, involving inverse covariances, which yields well-defined quantities. In other words, the Kalman filter and associated smoothing algorithms, in principle, have no difficulty in dealing with perfect information, corresponding to singularity of error covariances, while the information filter and its smoothing counterparts have no problem in dealing with a complete lack of information, corresponding to singularity of the *inverses* of error covariances.

However, what happens if we may have a complete lack of information about part of the state and perfect information about another part, so that neither the error covariance nor its inverse may be well-defined? In addition, what if the relationships between unknowns x_k and observations do not have as obvious a sequential structure as that displayed by the lower bi-diagonal matrix in (3)? When and how can we determine recursive estimation structures for such problems, generalizing both Kalman filtering and optimal smoothing algorithms for linear stochastic systems. In this paper we answer these questions by analyzing a rather general linear estimation problem whose study enables us to expand the range of applicability of Kalman filtering techniques to systems which are far more diverse and general than the usual state-space models. Such systems include for example both 1-D and 2-D stochastic descriptor systems, where the class of 2-D descriptor systems that we consider contains as special cases the 2-D state-space models of Roesser [3] and Fornasini and Marchesini [4], and can be used to model 2-D stochastic nearest-neighbor models of the type considered in [5]. In addition, the multiscale stochastic modeling and estimation framework developed in [10, 11, 12] also falls within the class of systems captured in our formalism.

The general estimation formulation adopted here is strongly influenced by our earlier work on the filtering and smoothing of 1-D descriptor systems [6], [7], where a general and flexible maximum likelihood (ML) approach was employed to derive recursive estimation algorithms. In addition in [15] Chisci and Zappa independently developed a square root Kalman filter for essentially the same problem studied in [7]. The main feature of the ML approach, which was itself motivated by earlier work of Whittle [8], Chapter 11, and of Bierman [9] in the context of square-root Kalman filtering, is that no distinction is made between system dynamics and observations. Specifically, all dynamic relations and initial or boundary conditions are viewed as observations, i.e. as noisy constraints on the state variables. Given a stream of observations, all observations considered up to a certain point can be compressed in such a way that the ML estimates based on the original observations or their compressed version are the same. Furthermore, observations concerning variables that are no longer of interest can be discarded. This process of compressing past observations and discarding unneeded variables is extended and generalized here in several ways. First, and most importantly, we introduce the concept of an *xo-graph*, which provides a unifying perspective for recursive estimation as well as an extremely convenient visualization of the structure of general linear estimation problems which in turn can be exploited to determine

the structure of recursive estimation algorithms for the broad array of problems mentioned previously. Secondly, with the exception of [7], the previous work (e.g. in [6],[15]) on estimation for 1-D descriptor systems under singular covariance and/or information matrix conditions has focused on "causal" filtering, i.e. recursive estimation of the "current state" given "present and past" observations. In this paper we consider "noncausal" smoothing as well providing both a generalization and a conceptually and notationally far simpler solution of the smoothing problem first analyzed in [7].

In the next section we state the general linear estimation problem of interest here, introduce its xo-graph representation, and illustrate by example the rich set of problems that are captured in this framework. It turns out that recursive estimation algorithms can be derived only for the class of so-called *acyclic* xo-graphs, and, since all xo-graphs are not necessarily acyclic, in Section 3 we develop aggregation operations on xo-graphs that can be used to reduce any such graph to an acyclic one. As we will see, this reduction directly provides a recursive structure for an estimation problem by identifying a grouping and sequential ordering of both the variables to be estimated and the observations to be processed. The general form of this reduction is that of a *tree*, leading to a generalization of the estimation problems considered in [10, 11, 12] for multiresolution stochastic processes. In Section 4 we then introduce the core operations required for recursive filtering and smoothing by considering several basic facts about ML estimation. In particular it is shown that two operations, called reduction and extraction, can be employed to compress observations, and discard unneeded variables. These two operations are then employed to derive recursive and numerically robust filtering and smoothing algorithms for ML estimation problems represented by trees. The case of linear trees is first discussed in Section 5, and the merge operation necessary to handle arbitrary trees is discussed in Section 6.

2 XO-Graphs for Linear Estimation

The general problem of interest to us is that of estimating a set of vectors $X = \{x_i\ ,\ i \in I\}$ with $x_i \in \mathbf{R}^{n_i}$, based on all or part of the set of linear observations:

$$o_k: \quad z_k = \sum_j^I A_{kj} x_j + G_k u_k\ ,\ k \in K \tag{4}$$

where the u_k's are zero-mean, independent Gaussian vectors with covariance $E[u_k u_k^T] = I$. For this estimation problem to possess a nontrivial recursive structure the observations in (1) should couple together only a "local" set of the x_j. For example in (3) each observation involves only a single x_j or two successive values. In our general problem, what we focus on is the structure of the sets of nonzero A_{kj} by associating to it a special type of graph, called an ***xo-graph***. An xo-graph has two types of nodes: x nodes corresponding to the unknown vectors x_i, and o nodes corresponding to the observations o_k. Each measurement (4) is represented by a set of J_k arcs, where J_k is the number of values of j for which $A_{kj} \neq 0$ and where the node o_k is connected to node x_j if the matrix $A_{kj} \neq 0$, i.e. if the unknown vector x_j contributes to the observation o_k. An important property of xo-graphs is that to go from a given x-node to another x-node, we must go through an o-node, and vice-versa. Graphs with this property are called bipartite [1].

Let us illustrate these ideas with several examples, a first simple one that we will use to illustrate some aspects of our construction and several others that indicate the generality of this framework

Example 1 The vectors to be estimated and observations are given by

$$X = \{x_1, x_2, x_3, x_4, x_5\} \tag{5a}$$
$$O = \{o_1, o_2, o_3, o_4\} \tag{5b}$$

with

$$o_1 : \ z_1 = A_{11}x_1 + A_{12}x_2 + G_1u_1 \tag{6a}$$
$$o_2 : \ z_2 = A_{21}x_1 + A_{22}x_2 + A_{23}x_3 + G_2u_2 \tag{6b}$$
$$o_3 : \ z_3 = A_{33}x_3 + A_{34}x_4 + G_3u_3 \tag{6c}$$
$$o_4 : \ z_4 = A_{44}x_4 + A_{45}x_5 + G_4u_4 \ , \tag{6d}$$

and the corresponding xo-graph is shown in Figure 1.

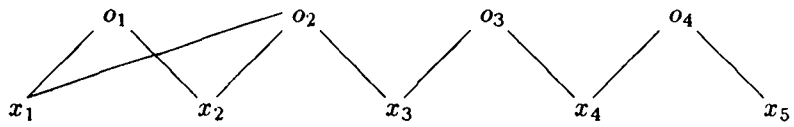

Figure 1: Xo-graph representing observations (3)

Note that by labeling each arc of the xo-graph with the corresponding A_{kj} matrix, and each node o_k with the observation vector z_k and matrix G_k, the given estimation problem can be totally represented by an xo-graph. The advantage of such a representation is that the structure of the xo-graph can be exploited to perform operations on the observations o_k aimed at estimating the unknown vectors x_j recursively. Also, a given xo-graph is said to be acyclic if the graph obtained by ignoring the x and o labels of the xo-graph does not contain any cycle. For example, the xo-graph of Figure 1 contains the x_1–o_1–x_2–o_2–x_1 cycle. In the next section we describe how to reduce such xo-graphs to acyclic ones. Since xo-graphs can always be decomposed into separate connected components, corresponding here to decoupled estimation problems, the xo-graphs that we shall consider when discussing recursive estimation algorithms are therefore *trees*, i.e. connected acyclic graphs.

Example 2 The Kalman filtering and smoothing problems for descriptor systems can be formulated in the form (4). To see this, consider the descriptor system

$$E_{k+1}x_{k+1} = A_kx_k + B_ku_k \ , \quad 0 \leq k \leq N-1 \tag{7}$$

with observations

$$y_{k+1} = C_{k+1}x_{k+1} + D_ku_k \ , \quad 0 \leq k \leq N-1 \ . \tag{8}$$

This model reduces to a standard state-space model when $E_k = I$, so that the descriptor estimation problem includes the corresponding problem for linear state-space models as a special case. In this model, it is assumed that the initial vector x_0 has zero mean and variance Π_0, and is independent of the noise u_k, which is assumed to be a zero-mean white Gaussian noise sequence with identity covariance. To transform the system (7)–(8) to the form (4), the main step is to view the system dynamics (7) as observations linking the state vectors x_k and x_{k+1}. Combining these observations with (8) yields

$$o_{k+1}: \quad z_{k+1} = \begin{pmatrix} 0 \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} -E_{k+1} \\ C_{k+1} \end{pmatrix} x_{k+1} + \begin{pmatrix} A_k \\ 0 \end{pmatrix} x_k + \begin{pmatrix} B_k \\ D_k \end{pmatrix} u_k \,, \tag{9}$$

with $0 \leq k \leq N-1$. The initial condition can similarly be transformed into an observation of the form

$$o_0: \quad 0 = x_0 + v_0 \,, \tag{10}$$

where v_0 is a zero-mean Gaussian vector independent of the noise u_k, with covariance matrix Π_0. The xo-graph corresponding to this estimation problem is shown in Figure 2. Clearly, it is connected and acyclic, so that it forms a tree. Since each node in this graph is connected to exactly two other nodes, we refer to this as a *linear* tree. In this context, the estimate of x_k based on all o's is the smoothed estimate of x_k, and its estimate based on the observations o_j such that $j \leq k$ is the filtered estimate.

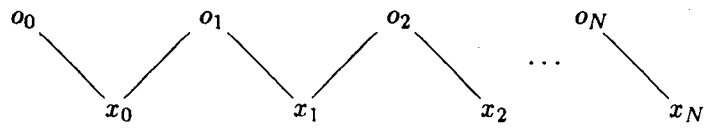

Figure 2: Xo-graph for descriptor systems.

Example 3 In Example 2, it was assumed that an initial condition was available for the descriptor system. For a two-point boundary value descriptor system (TPBVDS) [13]–[14], this initial condition is replaced by a boundary condition coupling the initial and final states x_0 and x_N, which can be modeled by an observation of the form

$$o_0: \quad z_N = A_{N0}x_0 + A_{NN}x_N + G_N u_N \,. \tag{11}$$

The xo-graph for this example contains the cycle x_0–o_1–x_1 ... o_N–x_N–o_0–x_0, as shown in Figure 3.

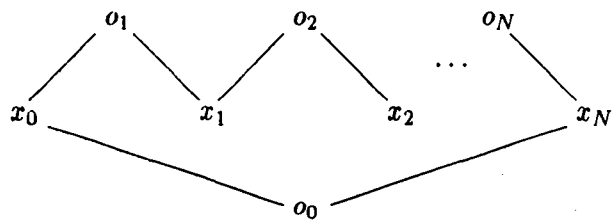

Figure 3: Xo-graph for boundary-value descriptor systems.

Example 4 In the preceding examples the index sets I (for the x_i) and K (for the z_k) are simply subsets of the integers. The framework we develop here can handle more general index sets including the 2-D index set used in 2-D systems. Many 2-D systems, such as those obtained by discretizing linear stochastic partial differential equations, can be described by a 2-D descriptor model of the form

$$A_0 x_{i+1j+1} + A_1 x_{ij+1} + A_2 x_{i+1j} + A_3 x_{ij} = B u_{ij} \,, \tag{12}$$

where u_{ij} is a 2-D white Gaussian noise sequence with unit covariance. This model includes as special cases the 2-D state-space models introduced by Roesser [3] and Fornasini and Marchesini [4], which correspond respectively to the choices

$$A_0 = 0 \quad A_1 = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \quad A_2 = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \tag{13}$$

and

$$A_0 = I \,. \tag{14}$$

It is also easy to verify that the 2-D nearest-neighbor stochastic models

$$z_{ij} = A_E z_{i-1j} + A_W z_{i+1j} + A_S z_{ij-1} + A_N z_{ij+1} + B u_{ij} \tag{15}$$

considered in [5] can be rewritten in the form (12) provided that we select

$$x_{ij}^T = [z_{ij}^T \; z_{i-1j-1}^T] \tag{16}$$

as partial state vector. For simplicity, it is assumed that the descriptor model (12) is defined over the rectangle $0 \leq i \leq N, 0 \leq j \leq M$, and the boundary conditions are of Dirichlet type, so that x_{ij} is known on the edges of the domain of definition. Then, given the observations

$$y_{ij} = Cx_{ij} + Du_{ij}\,, \tag{17}$$

we seek to find the ML estimate of x_{ij} based on all observations. To convert this estimation problem into the format (4), the dynamics (12) and observations (17) can be combined into a single observation o_{ij}, in the same manner as for the 1-D descriptor systems of Example 2. The resulting observation o_{ij} depends on x_{ij}, x_{i+1j}, x_{ij+1} and x_{i+1j+1}, so that the corresponding xo-graph has the structure shown in Figure 5. Clearly this graph contains many elementary cycles.

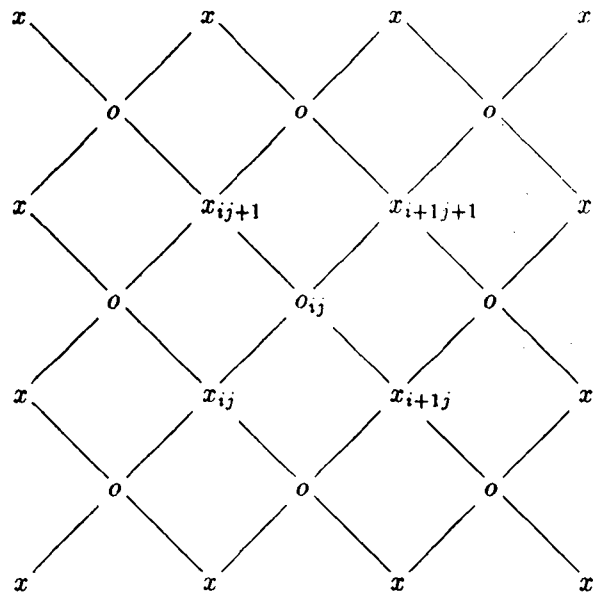

Figure 4: Xo-graph for 2-D descriptor systems.

Example 5 In [10, 11, 12] a class of recursive models on dyadic trees is introduced and studied as the basis for multiresolution modeling and processing of stochastic processes. In this setting the index set I consists of nodes on the tree corresponding to scale/translation pairs (m, n), where m denotes scale and n translational offset. As we move from one scale (say m) to the next finer one, $(m + 1)$, the number of points doubles and finer detail is added to the coarser description. The general form of such a coarse-to-fine recursion is

$$x_{i\alpha} = A_{i\alpha}x_i + B_{i\alpha}w_{i\alpha} \tag{18}$$

$$x_{i\beta} = A_{i\beta}x_i + B_{i\beta}w_{i\beta} \tag{19}$$

where if i denotes the pair (m, n), then $i\alpha = (m + 1, 2n)$ and $i\beta = (m + 1, 2n + 1)$ are the two descendents of i. If we also have observations

$$y_i = C_i x_i + v_i \tag{20}$$

the xo-graph has a national tree structure as depicted in Figure 5, where o_i consists of (18) - (20).

Finally, let us note that the examples given here also provide direct motivation for considering problems in which neither error covariances nor their inverses are well-defined. For example, consider the descriptor system (7) - (9) together with *separable boundary conditions*, i.e. boundary conditions as in (11) but which decouple into individual, partial conditions on each end point. Moreover, suppose that these boundary conditions are deterministic

$$V_o x_o = 0 \tag{21}$$

$$V_N x_N = 0 \tag{22}$$

where the ranks of V_o and V_N are each less than the corresponding dimensions of x_o and x_N. As a point of reference the reader should think of the discretization of the second-order differential equation $\ddot{s}(t) = u(t)$, where $u(t)$ is white noise, with split boundary conditions such as $s(0) = 0$, $\dot{s}(T) = 0$; in this case the discretization would involve a two-dimensional state with one-dimensional boundary constraints on each end.

Consider then the recursive estimation of x_k based on (7), (8) and (21). In this case we immediately start with (21) from which x_0 is not estimable (since V_0 has rank $<$ dim X_0), so that the corresponding error covariance of X_0 cannot be defined, but, on the other hand, part of X_0 is known *perfectly*, so that the information matrix is not defined either. Moreover, if

$$\begin{pmatrix} E_{k+1} \\ C_{k+1} \end{pmatrix}$$

is not full column rank (as it may not be if (7) comes from the discretization of a higher-order differential equation), then x_k will not be recursively estimable based on $o_j, j \leq k$. Said another way, if we collect the boundary conditions (21), (22), dynamics (7), and observations (8) into one set of simultaneous equations as we did in (3), the matrix we obtain in this case is also lower bi-diagonal:

$$\begin{bmatrix} V_0 & 0 & 0 & \cdots & & \\ -A_0 & E_1 & 0 & & & \\ 0 & C_1 & 0 & & & \\ & & & \ddots & & \\ & & & & -A_N & E_N \\ & & & \cdots & 0 & C_N \\ & & & \cdots & 0 & V_N \end{bmatrix} \tag{23}$$

In this case, while the full matrix may in fact be full rank (so that each x_k *is* estimable based on the use of *all* observations, including those for $j > k$), the upper left-hand submatrices

may *not* be, so that the straightforward application of Gaussian elimination, as in standard Kalman filtering and smoothing fails.

While the preceding example may appear somewhat special, it actually is not. In fact, as will be made clear in the next section, the general two-point boundary value problem of Example 3 results in a singular recursive estimation problem, as do essentially *all* 2-D estimation problems. In [7, 15] approaches to dealing with the singularity are developed to provide recursive filtering algorithms for estimating x_k based on o_j, $j \leq k$ for descriptor systems, i.e. for systems characterized by matrices as in (23), and in [10] a recursive solution is also given for the smoothing problem for this case. The results in the remainder of this paper generalize these earlier ones considerably.

3 XO-Graph Reduction

In this section we introduce and illustrate the use of the two types of operations that can be employed to reduce an arbitrary xo-graph to an acyclic one.

A X-aggregation

This operation consists of combining several x-nodes to form a larger x-node. The effect of this operation on the matrices A_{kj} appearing in the observation relations (4) is straightforward. Consider the case of Example 1. Then, the operation consisting of aggregating the nodes x_1, x_2 and x_3 of Figure 1 into a larger node is equivalent to stacking the corresponding vectors into a single column vector

$$x(1:3) = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \tag{24}$$

and the observations can be expressed in terms of this new unknown vector as

$$o_1: \quad z_1 = [A_{11}\ A_{12}\ 0]x(1:3) + G_1u_1 \tag{25a}$$
$$o_2: \quad z_2 = [A_{21}\ A_{22}\ A_{23}]x(1:3) + G_2u_2 \tag{25b}$$
$$o_3: \quad z_3 = [0\ 0\ A_{33}]x(1:3) + A_{34}x_4 + G_3u_3 \tag{25c}$$
$$o_4: \quad z_4 = A_{44}x_4 + A_{45}x_5 + G_4u_4\,. \tag{25d}$$

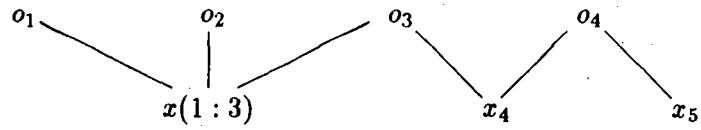

Figure 5: x-aggregation for Example 1.

The corresponding reduced xo-graph is depicted in Figure 6.

B O-aggregation

This operation is similar to x-aggregation except that we now combine several o nodes to form a single observation. This is accomplished by stacking the z vectors corresponding to the observations that we want to aggregate. Consider for example the xo-graph of Figure 1, and suppose that we want to aggregate the nodes o_1 and o_2. Introducing the stacked vector

$$z(1:2) = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \tag{26}$$

the observations o_1 and o_2 can be combined into a single observation

$$o(1:2): \quad \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} A_{11} \\ A_{31} \end{pmatrix} x_1 + \begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix} x_2 + \begin{pmatrix} 0 \\ A_{23} \end{pmatrix} x_3 + \begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}. \tag{27}$$

C Reduction to acyclic form

The two aggregation operations introduced above can be used to reduce an arbitrary xo-graph into an acyclic one. In general this can be done in a number of ways, resulting in different levels of aggregation and graph structures. The simplest way of accomplishing this objective consists of aggregating all x nodes and all o nodes together, which results in a trivial graph with a single x–o arc. This reduction technique is of course not very interesting since it destroys completely the dependency structure of the original xo-graph. A more sensible approach consists in minimizing the number of aggregation operations needed to reduce the given xo-graph into an acyclic one. To see how this approach works out in practice, consider the xo-graph of Figure 1. It turns out that this graph can be made acyclic by performing a single x-aggregation for the nodes x_1 and x_2, as shown in Figure 7.

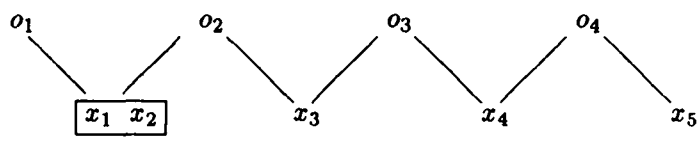

Figure 6: Reduced acyclic xo-graph for Example 1.

For the TPBVDS estimation problem represented by the xo-graph of Figure 3 the situation is slightly more complicated. To reduce this xo-graph to acyclic form, we can first

aggregate the state vectors x_i and x_{N-i} for $i = 0, 1, \ldots$. This yields the xo-graph of Figure 8.

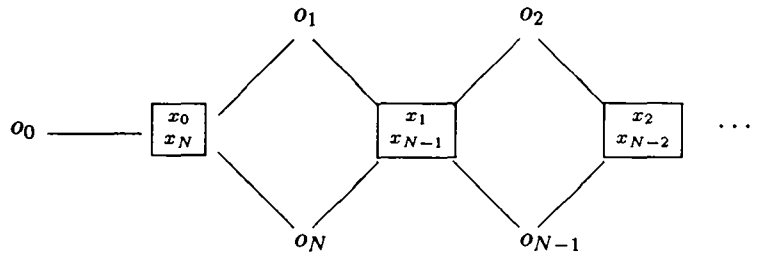

Figure 7: X-reduction of the TPBVDS xo-graph.

To eliminate the remaining cycles from the graph of Figure 8, we can aggregate the observations o_i and $o_{N-(i-1)}$ for $i = 1, 2, \ldots$, which gives the acyclic xo-graph of Figure 9.

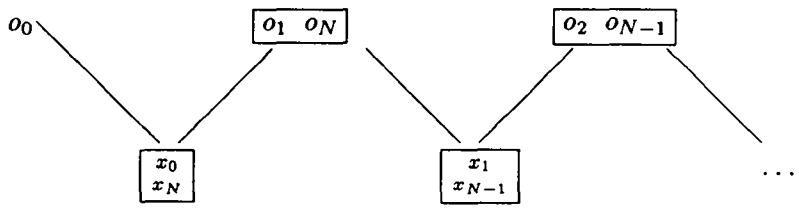

Figure 8: O-reduction of the TPBVDS xo-graph.

Note that the first observation o_0 in Figure 9 is the boundary condition (11), which generally provides an incomplete observation of (x_0, x_N) (e.g. for the example of a discretization of $\ddot{s}(t) = u(t)$, the boundary condition provides two constraints on the (discrete approximations to the) four variables $s(0), \dot{s}(0), s(T), \dot{s}(T)$. Thus recursive estimation in this case obviously involves dealing with singularity.

The xo-graph of Figure 4, which corresponds to the 2-D descriptor estimation problem of Example 4 is slightly more difficult to reduce. One reduction technique that leads to an acyclic xo-graph consists in organizing the observations o_{ij} and partial states x_{ij} into concentric regions, as shown in Figure 10. This means that the corresponding recursive estimation algorithms for estimating x_{ij} from the given model and observations will operate either outwards from the center of the observation region toward the edges, or inwards from the edges toward the center. However, this scheme is not the only one that can be used to reduce the given graph to an acyclic one. For example, by organizing the o_{ij}'s and x_{ij}'s columnwise or rowwise, it is possible to process the data from left to right or right to left, or from top to bottom, and vice-versa.

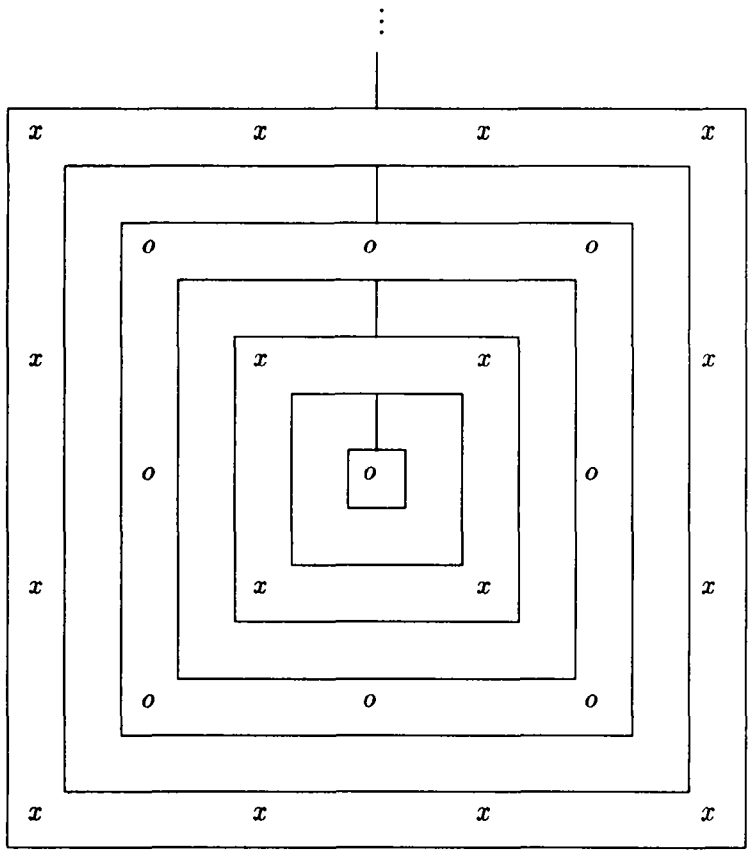

Figure 9: Reduced xo-graph for 2D descriptor estimation.

An interesting feature of the xo-trees obtained in Figures 9 and 10 for the TPBVDS and 2-D descriptor estimation problems, respectively, is that they are *linear*, i.e. they contain only one branch. A consequence of this property is that the recursive filtering and smoothing algorithms for standard Gauss-Markov state-space models will be applicable with only minor modifications to the above problems, since they correspond to the same xo-tree. In constrast, more general trees, such as the dyadic tree of Figure 5 require an interesting variation on the standard Kalman filter, described for the multiresolution model of Example 5 in [10, 11, 12] and generalized in Section 6. One property of these algorithms on trees is that they have an inherently parallel structure, as the processing on disjoint subtrees can be carried out in parallel. Indeed it is also possible to use aggregation to construct a tree structure for the 2-D problem of Figure 4. Specifically in this representation we first aggregate together the x_{ij} and o_{ij} along the central column of the 2-D array of Figure 4, yielding the nodes at the top of the tree and dividing the array into two halves. Each of these is then subdivided by aggregating the x_{ij} and o_{ij} along the central row of each half. Repeating this procedure on each of the disjoint rectangular regions produced in the preceding stage yields a tree structure that is used in [23] to develop multiresolution, pyramidal models for Markov random fields. In addition, as discussed in [24], the estimation

algorithm of Section 6 when applied to this example (in the nonsingular case) is very closely related to nested dissection methods [25] for solving system of linear equations.

4 Recursive Maximum Likelihood Estimation

Let x be an unknown vector in $\mathbf{R}^n$. Consider the problem of finding the maximum likelihood (ML) estimate of x given the observation

$$z = Ax + Gu \tag{28}$$

where z belongs to $\mathbf{R}^p$, and u is a zero-mean Gaussian vector in $\mathbf{R}^m$, with covariance matrix I_m. As we have discussed in Section 1, the problems motivating this paper require investigating a very general form of this ML problem. The static version of this problem has been analyzed by Campbell and Meyer in [26], and several aspects critical to recursive estimation have been developed under particular special conditions in [6, 7, 15]. In this section we generalize these ideas to the most general version of (28). First, as in [6], we do not assume that the covariance GG^T of the noise Gu is invertible. In other words, we allow the possibility that some measurements may be *perfect*. As we have pointed out, this extension is motivated by the fact that if we view the dynamics of a stochastic linear system as measurements, some of the dynamic relations for the system may not be affected by noise, and will therefore specify perfect, i.e. noiseless, measurements for some of the system variables. Another novel feature is that, unlike [6], we do not assume that the matrix A has full column rank. This means that all components of the vector x may not be estimable from the measurement (28). The need to consider ML estimation in such generality stems from the observation that the objective of recursive estimation is to incorporate progressively more information about a given system. In this context, although the final ML estimation problem may be well-posed, i.e. x may be estimable given all available measurements, this is not necessarily the case for intermediate estimation problems based only on a subset of measurements. For example, if we consider the descriptor system of Example 2, x_k may not be estimable based on the initial conditions and descriptor dynamics (9), since they do not necessarily specify a well-posed system, but it may become estimable after the observations (10) are included. Further, as discussed in [7], this situation is the rule rather than the exception for 2-D problems.

The ML estimation problem requires maximizing the probability density

$$p(u) = \frac{1}{(2\pi)^{m/2}} \exp -\frac{1}{2} u^T u \tag{29}$$

or equivalently, minimizing the quadratic cost $J(u) = u^T u/2$, under the constraint (29). It is shown in [6] that ML estimates can be obtained as follows.

Theorem 1 *$\hat{x}$ is a ML estimate of x if and only if for some λ, $\hat{x}$ satisfies*

$$\begin{pmatrix} GG^T & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \hat{x} \end{pmatrix} = \begin{pmatrix} z \\ 0 \end{pmatrix}. \tag{30}$$

The estimation error $\tilde{x} = x - \hat{x}$ *corresponding to such an estimate obeys*

$$\begin{pmatrix} GG^T & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ -\tilde{x} \end{pmatrix} = \begin{pmatrix} Gu \\ 0 \end{pmatrix}, \tag{31}$$

so that the bias vectors $m_\lambda = E[\lambda]$ *and* $b = x - E[\hat{x}]$ *satisfy*

$$\begin{pmatrix} GG^T & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} m_\lambda \\ -b \end{pmatrix} = 0\,. \tag{32}$$

Furthermore the error covariance matrix

$$\mathcal{P} \triangleq \begin{pmatrix} P_{\lambda\lambda} & -P_{\lambda\tilde{x}} \\ -P_{\tilde{x}\lambda} & P_{\tilde{x}\tilde{x}} \end{pmatrix} \tag{33a}$$

of the vector $[\lambda^T \; -\tilde{x}^T]^T$ *obeys*

$$\begin{pmatrix} GG^T & A \\ A^T & 0 \end{pmatrix} \mathcal{P} \begin{pmatrix} GG^T & A \\ A^T & 0 \end{pmatrix} = \begin{pmatrix} GG^T & 0 \\ 0 & 0 \end{pmatrix}. \tag{33b}$$

The equation (30) not only characterizes the ML estimates $\hat{x}$, but also provides a compatibility test for the measurement vector z, i.e. a test that can be employed to determine whether z originates from a measurement of the form (28). The vector z is compatible if and only if

$$\begin{pmatrix} z \\ 0 \end{pmatrix} \in \mathrm{Im} \begin{pmatrix} GG^T & A \\ A^T & 0 \end{pmatrix}. \tag{34}$$

Clearly, when z is compatible, the vector x admits at least one ML estimate, but this ML estimate is not necessarily unique.

A Reduction

A key aspect of recursive estimation in general is that we recursively compute and propagate a reduced or compressed version of the information collected – e.g. in standard causal Kalman filtering we propagate the estimate of the current state of the system. The generalization of this concept that we need for the general recursive ML estimation approach developed here is that of propagating a *sufficient statistic*, i.e. a compressed version of the collected information that is statistically equivalent to the original measurements for the purposes of recursive estimation. More precisely we shall freely replace a set of measurements for x by another, usually smaller, set of measurements that would yield the same family of ML estimates. This replacement procedure is justified by the introduction of the following concept.

Definition 1 *Two observations*

$$o_1: \; z_1 = A_1 x_1 + G_1 u_1 \tag{35a}$$

$$o_2: \; z_2 = A_2 x_1 + G_2 u_2 \tag{35b}$$

are said to be equivalent if for any other observation

$$o_3: \quad z_3 = A_3 x_1 + A_4 x_2 + G_3 u_3 \,, \tag{36}$$

with u_3 independent of u_1 and u_2, the set of ML estimates $(\hat{x}_1, \hat{x}_2)$ based on o_1 and o_3 is identical to the set of ML estimates $(\hat{x}_1, \hat{x}_2)$ based on o_2 and o_3.

Thus, two sets of observations are equivalent if they provide the same "information" about x. The idea of replacing a set of measurements by another containing the same information is not new and has been used informally in much of the recursive ML estimation and square-root Kalman filtering literature [9]. A notion of equivalence similar to the one introduced here was proposed recently in [15]. The version that we consider is slightly more general, since given two sets of measurements for the vector x_1, in order for these two measurements to be equivalent, we require not only that they should yield the same ML estimates of x_1, but also that they should be equivalent in terms of estimating any other vector x_2 for which additional measurements coupling x_1 and x_2 can be obtained.

To illustrate the concept of equivalence, we now describe a simple mechanism that can be employed to reduce an observation of a certain size to an equivalent observation of lower dimension.

Lemma 1 *Consider two observations*

$$o_1: \quad z_1 = A_1 x + G_1 u_1 \tag{37a}$$

$$o_2: \quad z_2 = G_2 u_2 \,, \tag{37b}$$

where u_1 and u_2 are two independent zero-mean Gaussian vectors with unit variance. Then the observation $o = o_1 \oplus o_2$ obtained by combining o_1 and o_2 is equivalent to o_1 only.

This result states the rather obvious fact that observations that do not involve x and which are statistically independent of the other observations can be removed.

Proof: Consider a third observation

$$o_3: \quad z_3 = A_2 x + A_3 y + G_3 u_3 \,, \tag{38}$$

where u_3 is independent of u_1 and u_3, and y is an unknown vector. To prove that o and o_1 are equivalent, we must show that the ML estimates $\hat{x}$ and $\hat{y}$ based on (37a)–(37b) and (38) are the same as those based on (37a) and (38). The ML estimates based on (37a)–(37b) and (38) satisfy

$$\begin{pmatrix} G_1 G_1^T & 0 & 0 & A_1 & 0 \\ 0 & G_2 G_2^T & 0 & 0 & 0 \\ 0 & 0 & G_3 G_3^T & A_2 & A_3 \\ A_1^T & 0 & A_2^T & 0 & 0 \\ 0 & 0 & A_3^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ 0 \\ 0 \end{pmatrix} \tag{39}$$

for some vectors λ_1, λ_2 and λ_3, whereas the estimates based on (37a) and (38) obey

$$\begin{pmatrix} G_1G_1^T & 0 & A_1 & 0 \\ 0 & G_3G_3^T & A_2 & A_3 \\ A_1^T & A_2^T & 0 & 0 \\ 0 & A_3^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_3 \\ \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} z_1 \\ z_3 \\ 0 \\ 0 \end{pmatrix} . \tag{40}$$

Since (40) is obtained by deleting the second block row and column of (39), all the solutions of (39) also satisfy (40). Conversely, suppose that λ_1, λ_3, $\hat{x}$ and $\hat{y}$ obey (40), and let z_2 be a measurement of the form (37b), so that it lies in the column space of G_2. Since the column spaces of G_2 and $G_2G_2^T$ are identical, the equation

$$G_2G_2^T\lambda_2 = z_2 \tag{41}$$

admits a solution, so that (39) is satisfied. This implies that o_1 is equivalent to o. 2

Among all equivalent measurements of a vector x, we can then ask ourselves whether it is possible to construct one which provides the most compact representation for the information contained in the measurement vector we are given, i.e. a minimal sufficient statistic. The feature that determines whether a measurement has been maximally compressed is as follows.

Definition 2 *The observation*

$$o: \quad z = Ax + Gu \tag{42}$$

is called reduced if A has full row rank.

To gain some intuition about the information contained in a reduced observation, consider the case of a reduced measurement where x is estimable, i.e. such that A has full column rank. Then the ML estimate of x is given by $\hat{x} = A^{-1}z$, and without loss of information we can premultiply (42) by A^{-1}, which yields

$$\hat{x} = x + \tilde{G}u \tag{43}$$

where $\tilde{G} = A^{-1}G$ and $P_{\tilde{x}\tilde{x}} = \tilde{G}\tilde{G}^T$ is the ML error covariance. Thus, in this case, a reduced observation just encodes the ML estimate of x and its error covariance. More generally, when x is not estimable, the reduced observation (42) can be viewed as encoding the ML estimate and error variance for the estimable part $x' = Ax$ of x.

The introduction of the concept of reduced observation is justified by the following result.

Theorem 2 *Every observation admits an equivalent reduced observation. Furthermore, if*

$$o_1: \quad z_1 = A_1x + G_1u_1 \tag{44a}$$

$$o_2: \quad z_2 = A_2x + G_2u_2 \tag{44b}$$

are two equivalent reduced observations, there exists an invertible matrix T such that

$$A_1 = TA_2 \tag{45a}$$

$$G_1G_1^T = TG_2G_2^TT^T . \tag{45b}$$

The relations (36) show that the reduced observation corresponding to a given measurement is unique up to left multiplication by an invertible matrix.

Proof: The proof of Theorem 2 proceeds in two stages. First, given an arbitrary observation, we show that it can be brought to the form (37a)–(37b) where the matrix A_1 has full row rank. Using Lemma 1, this implies that the given observation admits an equivalent reduced observation. Then, in a second stage, it is proved that two equivalent reduced observations must be related through (36).

Consider an arbitrary observation of the form (28). By performing a singular value decomposition of G, we can find orthonormal matrices U and V such that

$$VGU = \begin{pmatrix} G_1 & 0 \\ 0 & 0 \end{pmatrix} . \tag{46}$$

where the matrix G_1 is invertible. Premultiplying (28) by

$$S = \begin{pmatrix} G_1^{-1} & 0 \\ 0 & I \end{pmatrix} V , \tag{47}$$

and denoting

$$\begin{pmatrix} q \\ p \end{pmatrix} = Sz \qquad \begin{pmatrix} Q \\ P \end{pmatrix} = SA \tag{48a}$$

$$\begin{pmatrix} m \\ n \end{pmatrix} = U^T u , \tag{48b}$$

the given observation can be decomposed as

$$q = Qx + n \qquad p = Px , \tag{49}$$

where the covariance of the noise n is the identity matrix. The expression (49) corresponds to a decomposition of the original observation into a part where the noise is nonsingular, and a perfect observation. Then, by performing a QR decomposition of P, we can premultiply p by an orthonormal matrix T such that

$$Tp = \begin{pmatrix} p_1 \\ 0 \end{pmatrix} = \begin{pmatrix} P_1 \\ 0 \end{pmatrix} x , \tag{50}$$

where P_1 has full row rank. Next, we project the rows of Q onto the space spanned by the rows of P_1, so that if K denotes the corresponding projection matrix, the rows of $\tilde{Q} = Q - KP_1$ are orthogonal to the rows of P_1. This yields a modified measurement of the form

$$\tilde{q} = q - Kp_1 = \tilde{Q}x + n . \tag{51}$$

Finally, by performing a QR decomposition of $\tilde{Q}$, we can find an orthonormal matrix W such that

$$W\tilde{q} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} = \begin{pmatrix} R_1 \\ 0 \end{pmatrix} + w , \tag{52}$$

where R_1 has full row rank and the covariance of the noise

$$Wn = w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} , \tag{53}$$

is the identity matrix, since the covariance of n was also identity. This implies that w_1 and w_2 are independent. Thus, by a sequence of reversible operations, we have transformed the original observation to the form

$$\begin{pmatrix} p_1 \\ r_1 \end{pmatrix} = \begin{pmatrix} P_1 \\ R_1 \end{pmatrix} x + \begin{pmatrix} 0 \\ w_1 \end{pmatrix} \tag{54a}$$

$$r_2 = w_2 \tag{54b}$$

which is in the form (37a)–(37b), where by construction, the matrix $[P_1^T \; R_1^T]^T$ has full row rank. Thus, by Lemma 1, the original observation is equivalent to (54a), which is reduced.

To complete the proof of Theorem 2, we need to prove that if o_1 and o_2 are two equivalent reduced observations of the form (44a) and (44b), they are necessarily related through (36). To see this, note that the ML estimates of x based on o_i with $i = 1, 2$ satisfy

$$\begin{pmatrix} G_1 G_i^T & A_i \\ A_i^T & 0 \end{pmatrix} \begin{pmatrix} \lambda_i \\ \hat{x} \end{pmatrix} = \begin{pmatrix} z_i \\ 0 \end{pmatrix}. \tag{55}$$

Since the observation o_i is reduced A_i^T has full column rank, so that $\lambda_i = 0$, and (55) reduces to

$$z_i = A_i \hat{x} \tag{56}$$

for $i = 1, 2$. In order for the solutions of equation (55) to be the same for $i = 1, 2$, A_1 and A_2 must have the the same right null space. Since A_1 and A_2 have full row rank, this means that they must have the same reduced row echelon form, so that there exists an invertible matrix T such that (45a) holds. In addition, to ensure that the solutions of (55) are the same for $i = 1, 2$, we must also have $z_1 = Tz_2$, which in combination with (45a) implies (45b). 2

B Extraction

Another operation that is needed in deriving recursive estimation algorithms involves discarding unneeded variables that are no longer of interest. Equivalently, given a measurement, we want to be able to extract a submeasurement concerning only the variables in which we are still interested. For example in standard Kalman filtering, the prediction step in fact corresponds to an extraction of relevant information about x_{k+1} from previous measurements and current dynamics and the dropping of the estimate of x_k from the set of statistics to be updated when the next measurement is to be incorporated. This operation, which will be called *extraction*, was first described in [6], [7] and was also introduced in [15] in the context of square-root Kalman filtering for descriptor systems. The main difficulty in performing an extraction is that we want to ensure that we are not throwing away any useful information concerning the variables that are of interest. This requirement can be expressed as follows.

Definition 3 *The observation*

$$o_1 : \; z_1 = Ax_1 + G_1 u_1 \tag{57}$$

is said to be an <u>*extraction*</u> *of the vector* x_1 *from the measurement*

$$o_0: \quad z_0 = A_0 x_0 + A_1 x_1 + G_0 u_0 \tag{58}$$

if for all observations

$$o_2: \quad z_2 = A_2 x_1 + A_3 x_2 + G_2 u_2\,, \tag{59}$$

with u_2 *independent of* u_0 *and* u_1, *the set of ML estimates* $\hat{x}_1$, $\hat{x}_2$ *based on* o_1 *and* o_2 *is identical to the set of ML estimates* $\hat{x}_1$, $\hat{x}_2$ *based on* o_0 *and* o_2.

The following result provides a general mechanism for performing extractions.

Theorem 3 *The observation*

$$o_1: \quad Lz = LA_1 x_1 + LG_1 u \tag{60}$$

is an extraction of x_1 *from*

$$o: \quad z = A_0 x_0 + A_1 x_1 + G_1 u \tag{61}$$

if L *is a basis of the left null space of* A_0, *i.e., it is a matrix of maximum rank such that* $LA_0 = 0$.

Proof: First observe that to construct L we need only to perform a QR factorization of A_0. Specifically, let Q be an orthonormal matrix such that

$$\begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} A_0 = \begin{pmatrix} M_1 \\ 0 \end{pmatrix}, \tag{62}$$

where M_1 has full row rank. Then $L = Q_2$. To prove that (60) is an extraction of (61), note that the ML estimates of x_1 and x_2 based on (61) and (59) must satisfy

$$\begin{pmatrix} G_1 G_1^T & 0 & A_0 & A_1 & 0 \\ 0 & G_2 G_2^T & 0 & A_2 & A_3 \\ A_0^T & 0 & 0 & 0 & 0 \\ A_1^T & A_2^T & 0 & 0 & 0 \\ 0 & A_3^T & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \hat{x}_0 \\ \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{63}$$

for some λ_1, λ_2 and $\hat{x}_0$, whereas the ML estimates of x_1 and x_2 based on the extracted measurement (60) and (59) obey

$$\begin{pmatrix} LG_1 G_1^T L^T & 0 & LA_1 & 0 \\ 0 & G_2 G_2^T & A_2 & A_3 \\ A_1^T L^T & A_2^T & 0 & 0 \\ 0 & A_3^T & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_e \\ \lambda_2 \\ \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \begin{pmatrix} Lz_1 \\ z_2 \\ 0 \\ 0 \end{pmatrix}. \tag{64}$$

Note that the second and last block rows of (63) and (64) are identical.

To show that the ML estimates $\hat{x}_1$ and $\hat{x}_2$ which solve (63) and (64) are the same, assume first that λ_1, λ_2, $\hat{x}_0$, $\hat{x}_1$ and $\hat{x}_2$ satisfy (63). The third block row

$$A_0^T \lambda_1 = 0 \tag{65}$$

of this equation implies that λ_1 is in the left null space of A_0, and since L is a basis of this null space, there exists a vector λ_e such that

$$\lambda_1 = L\lambda_e \,. \tag{66}$$

Substituting this relation inside the fourth block row of (63) yields the third block row of (64), and multiplying the first block row of (63) by L and taking into account $LA_0 = 0$, we obtain the first block row of (64). Thus, the vectors λ_e, λ_2, $\hat{x}_1$ and $\hat{x}_2$ satisfy (64).

Conversely, let the vectors λ_e, λ_2, $\hat{x}_1$ and $\hat{x}_2$ obey (64). Then define the vector λ_1 through (66). This implies that the third block row of (64) is the same as the fourth block row of (63), and since L is a basis of the left null space of A_0, we have

$$A_0^T \lambda_1 = 0 \tag{67}$$

so that the third block row of (65) is satisfied. Finally, consider the first block row

$$L(G_1 G_1^T \lambda_1 + A_1 \hat{x}_1 - z_1) = 0 \tag{68}$$

of (64). This implies that the vector

$$a \triangleq G_1 G_1^T \lambda_1 + A_1 \hat{x}_1 - z_1 \tag{69}$$

is orthogonal to the left null space of A_0, so that it must be in its column space, i.e. we can find $\hat{x}_0$ such that $a = -A_0 \hat{x}_0$, which implies that the first block row of (63) is satisfied. This shows that the ML estimates $\hat{x}_1$ and $\hat{x}_2$ based on (59) and (61) are the same as those based on (59) and (60), so that (60) is an extraction of (61). 2

C Recursive Estimation

We have now all the elements necessary to develop recursive ML estimation algorithms. The algorithms considered below will be based entirely on the following *elementary operations*:

(i) Aggregating two observations: $o = o_1 \oplus o_2$

(ii) Reducing an observation o: $o_r = R\{o\}$.

(iii) Extracting a vector x from an observation o: $o_x = X_x\{o\}$.

(iv) The equivalence of two observations o_1 and o_2: $o_1 \equiv o_2$. when o_1 and o_2 are both reduced, this means that they can be obtained from each other by left multiplication by an invertible matrix.

Note that since the reduction and extraction operations described above rely on numerically stable techniques such as the singular value decomposition, QR factorizations, and orthonormal projections, all algorithms obtained by combining such operations can be implemented in a numerically reliable manner. In fact, is easy to verify that the square-root Kalman filtering algorithms [9], [2], [15] precisely on operations of this type. The recursive estimation algorithms described in the next two sections rely on the following result, which is a direct consequence of Theorems 2 and 3.

Theorem 4 *Consider the observations*

$$o_1: \quad z_1 = A_1x_1 + A_2x_2 + G_1u_1 \tag{70a}$$

$$o_2: \quad z_2 = A_3x_2 + A_4x_3 + G_2u_2\,. \tag{70b}$$

Then

$$R\{X_{x_3}\{o_1 \oplus o_2\}\} \equiv R\{X_{x_3}\{R\{X_{x_2}\{o_1\}\} \oplus o_2\}\}\,. \tag{71}$$

Thus, we can either perform extraction and reduction operations in one step by working with all observations together, or we can perform these operations recursively, as more observations become available. That is, we can first process (70a), keeping only a reduced form of the information concerning x_2 contained in o_1 and then can combine this with (70b), allowing use to repeat the process by extracting x_3 and again transforming this into reduced form.

5 Estimation on Linear Trees

In this section, we consider estimation problems corresponding to linear xo-trees, which are trees with a single branch. The main motivation for considering such trees arises from the observation that standard state-space estimation problems have exactly this tree structure. Thus, provided that they are expressed abstractly in terms of the elementary operations introduced in the previous section, all the standard Kalman filtering and smoothing algorithms are applicable to this family of trees. Another motivation for considering such trees is that many problems that do not obviously have such a structure can, through the use of the x- and o-aggregation operations introduced in Section 2, be reduced to linear trees. This is the case for example of the TPBVDS and 2-D descriptor estimation problems whose reduced graphs appear in Figures 8 and 9, respectively.

Consider the estimation problem

$$o_0: \quad z_0 = E_0x_0 + G_0u_0 \tag{72a}$$

$$o_k: \quad z_k = E_kx_k + A_{k-1}x_{k-1} + G_{k-1}u_{k-1}\,, \quad 1 \le k \le N \tag{72b}$$

$$o_{N+1} \quad z_{N+1} = A_Nx_N + G_{N+1}u_{N+1} \tag{72c}$$

whose xo-graph is the linear tree shown in Figure 11.

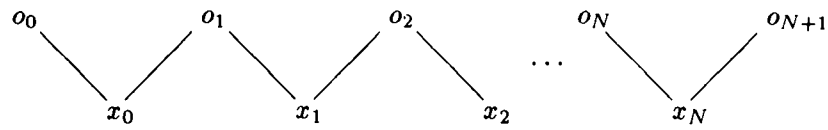

Figure 10: Linear tree.

A Filtering

The Kalman filtering algorithm for this problem can be described as follows. Let $\hat{o}_k^f$ be the observation obtained by extracting x_k from the combination of all observations o_j with $0 \leq j \leq k$, and then reducing the resulting extracted observation, i.e.

$$\hat{o}_k^f \triangleq R\{X_{x_k}\{\oplus_{j=0}^k o_j\}\} . \tag{73}$$

Then, Theorem 4 implies that $\hat{o}_k^f$ satisfies the recursion

$$\hat{o}_0^f = R\{o_0\} \tag{74a}$$

$$\hat{o}_{k+1}^f = R\{X_{x_{k+1}}\{\hat{o}_k^f \oplus o_{k+1}\}\} \tag{74b}$$

for $0 \leq k \leq N-1$, which is the abstract form of the forward Kalman filter. In these recursions, the observation $\hat{o}_k^f$ represents a complete summary of the information about x_k contained in the past observations o_j, $0 \leq j \leq k$. To the observation $\hat{o}_k^f$, we can then associate a family of ML estimates $\hat{x}_k^f$ of x_k based on the past observations, where the estimate $\hat{x}_k^f$ is unique only if x_k is estimable from the observations o_j with $0 \leq j \leq k$. This is ensured in particular if the matrices E_k have full rank for all k. When x_k is estimable, the observation $\hat{o}_k^f$ can be represented as

$$\hat{o}_k : \ \hat{x}_k^f = x_k + P_k^{f\,1/2} v_k , \tag{75}$$

where $P_k^{f\,1/2}$ is a square-root of the error covariance matrix P_k^f of the estimate $\hat{x}_k^f$, and v_k is a zero-mean Gaussian vector with unit variance.

Example 2, continued: The Kalman filtering problem for descriptor systems has been investigated in [16], [17], [6], [15]. When x_k is estimable from past observations, a general 3-block form for the optimum filter and its associated Riccati equation were obtained in [6], which includes also a complete analysis of the steady-state convergence of the optimum filter. The case when x_k is not estimable was subsequently examined in [7], [18], [15].

For descriptor systems, it is shown in Section 1 that the observations (63) take the form (9). Then, the recursive filtering scheme (65), when applied to these observations, reduces exactly to the square-root algorithm of [15]. It is also easy to verify that when x_k is estimable, the recursions (65) yield the 3-block Kalman filter of [6]. To see this, note that the first step of the recursion (74b) requires extracting x_{k+1} from the combination of $\hat{o}_k$ and observation (9). But, when x_k is estimable, $\hat{o}_k^f$ can be expressed in the form (75), and

the extraction step consists in backsubstituting $x_k = \hat{x}_k^f - P_k^{f1/2} v_k$ inside (9). This yields the observation

$$\begin{pmatrix} A\hat{x}_k^f \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} E_{k+1} \\ C_{k+1} \end{pmatrix} x_{k+1} + \begin{pmatrix} A_k P_k^{f1/2} & -B_k \\ 0 & D_k \end{pmatrix} \begin{pmatrix} v_k \\ u_k \end{pmatrix}, \tag{76}$$

where v_k and u_K are independent. From this observation, we see that x_{k+1} will be estimable provided the matrix $[E_{k+1}^T \; C_{k+1}^T]^T$ has full column rank, which is precisely the condition obtained in [6]. Then, the next step of the recursion (74b) requires reducing the observation (76). One way to achieve this is to apply the square-root procedure described in the proof of Theorem 2. However, the resulting filter has an implicit form. To obtain a closed-form expression for the filter, we need only to note that according to Theorem 1, the ML estimate $\hat{x}_{k+1}^f$ of x_{k+1} satisfies

$$\begin{pmatrix} A_k P_k^f A_k^T + Q_k & -S_k & E_{k+1} \\ -S_k^T & R_k & C_{k+1} \\ E_{k+1}^T & C_{k+1}^T & 0 \end{pmatrix} \begin{pmatrix} \xi_k \\ \lambda_k \\ \hat{x}_{k+1}^f \end{pmatrix} = \begin{pmatrix} A_k \hat{x}_k^f \\ y_{k+1} \\ 0 \end{pmatrix} \tag{77a}$$

with

$$Q_k = B_k B_k^T \quad S_k = B_k D_k^T \quad R_k = D_k D_k^T, \tag{77b}$$

from which we deduce the 3-block expression

$$\hat{x}_{k+1}^f = \begin{pmatrix} 0 & 0 & I \end{pmatrix} \begin{pmatrix} A_k P_k^f A_k^T + Q_k & -S_k & E_{k+1} \\ -S_k^T & R_k & C_{k+1} \\ E_{k+1}^T & C_{k+1}^T & 0 \end{pmatrix}^\dagger \begin{pmatrix} A_k \hat{x}_k^f \\ y_{k+1} \\ 0 \end{pmatrix} \tag{78}$$

which was obtained for the filter in [6]. Here $M^\dagger$ denotes the Moore-Penrose pseudo-inverse [19], p. 243 of a matrix M. Similarly, the expression (34a for the error covariance of a ML estimate yields the 3-block Riccati equation

$$P_{k+1}^f = -\begin{pmatrix} 0 & 0 & I \end{pmatrix} \begin{pmatrix} A_k P_k^f A_k^T + Q_k & -S_k & E_{k+1} \\ -S_k^T & R_k & C_{k+1} \\ E_{k+1}^T & C_{k+1}^T & 0 \end{pmatrix}^\dagger \begin{pmatrix} 0 \\ 0 \\ I \end{pmatrix}. \tag{79}$$

2

B Smoothing

With the exception of [7], [18], most treatments of recursive ML estimation for problems of the type considered here (or special cases thereof as in [6]) have dealt exclusively with problems of filtering. In this subsection we present the generalizations of the two-filter and double-sweep smoothing algorithms [20], [21], [22] for standard state-space models to the general ML estimation problem on linear trees. As we will see, while the two-filter generalization is straightforward, the double-sweep or Rauch-Tung-Striebel (RTS) algorithm has a subtle twist, due to the nature of the extraction and reduction processes, in order to deal with the fact that variables that are not estimable using past data may become estimable when future data is included as well.

To begin, we construct a backward Kalman filter which is the counterpart of the forward filter (65) in the sense that it starts from the other end of the tree and propagates in the opposite direction. Let $\hat{o}_k^b$ be the observation obtained by extracting the vector x_k from the combination of the observations o_j such that $k+1 \leq j \leq N+1$, and then reducing the resulting extracted observation, i.e.

$$\hat{o}_k^b = R\{X_{x_k}\{\oplus_{k+1}^{N+1} o_j\}\}\,. \tag{80}$$

Then, $\hat{o}_k^b$ can be computed recursively with the backward Kalman filter

$$\hat{o}_N^b = R\{o_{N+1}\} \tag{81a}$$
$$\hat{o}_k^b = R\{X_{x_k}\{\hat{o}_{k+1}^b \oplus o_{k+1}\}\}\,. \tag{81b}$$

Consider now the smoothed observation $\hat{o}_k^s$ obtained by extracting x_k from all observations, and then reducing the resulting extracted observation, so that

$$\hat{o}_k^s = R\{X_{x_k}\{\oplus_{j=0}^{N+1} o_j\}\}\,. \tag{82}$$

According to Theorem 4, $\hat{o}_k^s$ can be constructed by extracting x_k separately from the past and future observations, and reducing the resulting observations, which gives $\hat{o}_k^f$ and $\hat{o}_k^b$, and then reducing the combination of these two observations. Thus, we have

$$\hat{o}_k^s = R\{\hat{o}_k^f \oplus \hat{o}_k^b\}\,, \tag{83}$$

which is the *two-filter* smoothing formula for the given tree estimation problem.

It reduces to the usual two-filter smoothing formula when x_k is estimable separately from the past and future observations, and the corresponding covariance matrices P_k^f and P_k^b are positive definite. This can be verified by noting that under these assumptions, the observations $\hat{o}_k^f$ and $\hat{o}_k^b$ can be expressed as

$$\hat{o}_k^f : \; d_k^f \triangleq (P_k^f)^{-1/2}\hat{x}_k^f = (P_k^f)^{-1/2}x_k + u_k^f \tag{84a}$$
$$\hat{o}_k^b : \; d_k^b \triangleq (P_k^b)^{-1/2}\hat{x}_k^b = (P_k^b)^{-1/2}x_k + u_k^b\,, \tag{84b}$$

where u_k^f and u_k^b are two independent zero-mean Gaussian vectors with unit intensity. Then, when the reduction operation described in the proof of Theorem 2 is applied to the combination $\hat{o}_k^f \oplus \hat{o}_k^b$, it requires finding an orthonormal matrix T such that

$$T\begin{pmatrix} (P_k^f)^{-1/2} & d_k^f \\ (P_k^b)^{-1/2} & d_k^b \end{pmatrix} = \begin{pmatrix} (P_k^s)^{-1/2} & d_k^s \\ 0 & a_k \end{pmatrix}\,, \tag{85b}$$

where $\hat{x}_k^s$ and P_k^s denote the smoothed estimate of x_k and its error variance, $d_k^s = P^{s-1/2}\hat{x}_k^s$, and a_k is an arbitrary vector. Premultiplying (85b) by its transpose, and taking into account the orthogonality of T, yields

$$(P_k^s)^{-1} = (P_k^f)^{-1} + (P_k^b)^{-1} \tag{86a}$$
$$(P_k^s)^{-1}\hat{x}_k^s = (P_k^f)^{-1}\hat{x}_k^f + (P_k^b)^{-1}\hat{x}_k^b \tag{86b}$$

which are the usual two-filter smoothing relations.

Turning to the generalization of the RTS smoothing formula for a linear xo-tree, we find that we must be a bit careful in developing this result. In particular the usual RTS smoother for causal systems consists of a forward Kalman filter to process the data and a reverse sweep that processes the filtered estimates alone in order to produce the smoothed estimates. If one were to write this smoothing problem as a large, static estimation problem – i.e. as in (3) – we would find that this procedure corresponds simply to a Gaussian elimination step (the Kalman filter) on the block tridiagonal normal equations arising from (3) and a back substitution (the reverse sweep) to yield the smoothed estimate. In our more general problem of ML estimation on linear trees, we also have a tridiagonal set of normal equations, and the Kalman filtering reduction/extraction procedure described previously corresponds to the Gaussian elimination step with one significant difference, namely that because of the possibility that x_k may not be estimable based only on the past, *the filtering procedure given by (74a) may discard some measurements that are of no value for filtering but may be of use for smoothing.*

In particular, suppose that we have $\hat{o}_k^f$ as in (73) and suppose that some part of x_k is not estimable based on $\hat{o}_k^f$. More precisely, suppose that some part of $A_k x_k$ is not estimable. In this case the incorporation of o_{k+1} as in (72b) will include some nonestimable portion of x_k. For recursive *filtering*, however, at this point we are longer interested in x_k but rather in x_{k+1}, and the result is that the extraction process for x_{k+1} in (74b) will discard that portion of o_{k+1} that contains nonestimable parts of x_k. That is, at each step in the Gaussian elimination process we *discard* some of the equations, i.e. we ignore some measurements, which are not important for filtering. However, these discarded pieces of information may very well be of value for *smoothing* (e.g. if $E_{k+1}x_{k+1}$ is estimable based on future data, than o_{k+1} will provide useful information for the part of $A_k x_k$ not estimable based solely on past information).

The net result of all of this is that, unlike the standard causal case, the backward sweep of the general RTS algorithm will also involve the processing of at least a part of the raw data (corresponding to the discarded measurements). To see how to do this, we note that the culprit here is the coupling between x_k and x_{k+1} in o_{k+1}. Thus, let us collect *all* information about these two variables into three sets: the set prior to o_{k+1} (which involves x_k but not x_{k+1}), o_{k+1} itself, and the information subsequent to o_{k+1} (involving x_{k+1} but not x_k). That is, we observe that all the information about x_k and x_{k+1} contained in the observations o_j with $0 \leq j \leq N+1$ is also contained in the compressed observation $\hat{o}_k^f \oplus o_{k+1} \oplus \hat{o}_{k+1}^b$. Since we do not necessarily assume that x_k is separately estimable from the past or future observations alone, the forward and backward observations $\hat{o}_k^f$ and $\hat{o}_{k+1}^b$ can be assumed to take the general form

$$\hat{o}_k^f : \quad z_k^f = L_k^f x_k + G_k^f u_k^f \tag{87a}$$

$$\hat{o}_{k=1}^b : \quad z_{k+1}^b = L_{k+1}^b x_{k+1} + G_{k+1}^b u_{k+1}^b \tag{87b}$$

where u_k^f and u_{k+1}^b are independent with unit variance. Then, according to Theorem 1, the

smoothed estimates $\hat{x}_k^s$ and $\hat{x}_{k+1}^s$ satisfy the system

$$\begin{pmatrix} G_k^f G_k^{fT} & 0 & 0 & L_k^f & 0 \\ 0 & G_k G_k^T & 0 & A_k & E_{k+1} \\ 0 & 0 & G_{k+1}^b G_{k+1}^{bT} & 0 & L_{k+1}^b \\ L_k^{fT} & A_k^T & 0 & 0 & 0 \\ 0 & E_{k+1}^T & L_{k+1}^{bT} & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_f \\ \xi \\ \lambda_b \\ \hat{x}_k^s \\ \hat{x}_{k+1}^s \end{pmatrix} = \begin{pmatrix} z_k^f \\ z_{k+1} \\ z_{k+1}^b \\ 0 \\ 0 \end{pmatrix} \tag{88}$$

for some vectors λ_f, ξ and λ_b. Dropping the third and fifth block rows of (88) gives

$$\begin{pmatrix} G_k^f G_k^{fT} & 0 & L_k^f \\ 0 & G_k G_k^T & A_k \\ L_k^{fT} & A_k^T & 0 \end{pmatrix} \begin{pmatrix} \lambda_f \\ \xi \\ \hat{x}_k^s \end{pmatrix} = \begin{pmatrix} z_k^f \\ z_{k+1} - E_{k+1}\hat{x}_{k+1}^s \\ 0 \end{pmatrix} . \tag{89}$$

If we now assume that x_k is estimable from *all* observations, the relation (89) yields the Rauch-Tung-striebel smoothing recursion

$$\hat{x}_k^s = \begin{pmatrix} 0 & 0 & I \end{pmatrix} \begin{pmatrix} G_k^f G_k^{fT} & 0 & L_k^f \\ 0 & G_k G_k^T & A_k \\ L_k^{fT} & A_k^T & 0 \end{pmatrix}^\dagger \begin{pmatrix} z_k^f \\ z_{k+1} - E_{k+1}\hat{x}_{k+1}^s \\ 0 \end{pmatrix} , \tag{90}$$

where the smoothed estimate $\hat{x}_k^s$ is obtained by first propagating the Kalman filter (65) in the forward direction, which gives the observations $\hat{o}_k^f$, and then propagating (90) in the backwards direction, so that the whole tree is swept twice in opposite directions. In general, *both* of these sweeps use the original data z_k. However, by carefully organizing the observations (72b) in the standard causal case, it is straightforward to recover the usual RTS algorithm.

6 Arbitrary Trees

In [10, 11, 12] the estimation problem for multiscale processes on dyadic trees, as described by the model (18)-(20) of Example 5, is considered, and the generalization of the RTS algorithm is developed for this problem. In this section we consider the more general problem of possibly singular measurements on arbitrary trees, and we describe the extensions of both the two-filter and RTS algorithms to this setting. To begin this development, it is useful to observe that the forward and backward Kalman filters for linear trees were initialized at each extremity of the tree, and then combined all tree observations sequentially, in the order in which they were encountered. The same principle applies for arbitrary trees: the filtering and smoothing smoothing algorithms that we develop in this section rely on initializing a filter at each extremity of the tree, and then merging the outputs of different filters as we move inward from the extremities of the tree. The key additional operation needed to perform this task is the generalization of the *merge operation* introduced in [10, 11, 12], whereby the outputs of Kalman filters characterizing the observations contained in nonoverlapping subtrees are combined to yield estimates which now summarize the observations on the subtree formed by the union of the merging subtrees.

A Merge operation

Consider an o-node connected to the nodes x_j, with $1 \leq j \leq N$, as shown in Figure 12. Then, if $\mathcal{T}$ denotes the given tree, the tree $\mathcal{T} - \{o\}$ obtained by removing from the node o and the arcs connected to it from $\mathcal{T}$ can be partitioned into N subtrees $\mathcal{T}_j$, where a node belongs to the subtree $\mathcal{T}_j$ if the unique path connecting it to o passes through node x_j. Note that since o is only connected to the nodes x_j, $1 \leq j \leq N$, any path leading to o must necessarily go through one of these nodes. Let O be the observation obtained by o-aggregation of all observations contained in the tree $\mathcal{T}$. Let also O_j be the observation obtained by aggregating all observations of the subtree $\mathcal{T}_j$. The observation O can be decomposed as

$$O = O_1 \oplus \check{O}_1 \tag{91a}$$

$$\check{O}_1 = \left(\oplus_{j=2}^{N} O_j\right) \oplus o\,, \tag{91b}$$

where $\check{O}_1$ corresponds to the observation obtained by removing all observations contained in subtree $\mathcal{T}_1$ from O, or equivalently, by aggregating the observations O_j for $j \neq 1$ with the o-node observation. Then, let $\hat{o}_j$ be the observation obtained by extracting x_j from O_j, and reducing the resulting observation, so that

$$\hat{o}_j = R\{X_{x_j}\{O_j\}\}\,. \tag{92}$$

Similarly, let $\check{o}_1$ be the observation obtained by extracting x_1 from $\check{O}_1$ and reducing the resulting observation, i.e.

$$\check{o}_1 = R\{x_{x_1}\{\check{O}_1\}\} \tag{93}$$

Then the merge operation

$$\check{o}_1 = R\{X_{x_1}\{\left(\oplus_{j=2}^{N} \hat{o}_j\right) \oplus o\}\} \tag{94}$$

is the extension to arbitrary trees of the forward and backward Kalman filtering identities obtained in the previous section. It is a direct consequence of Theorem 4, and provides a mechanism for recursively processing the tree observations, starting from the extremities of the tree and moving inwards.

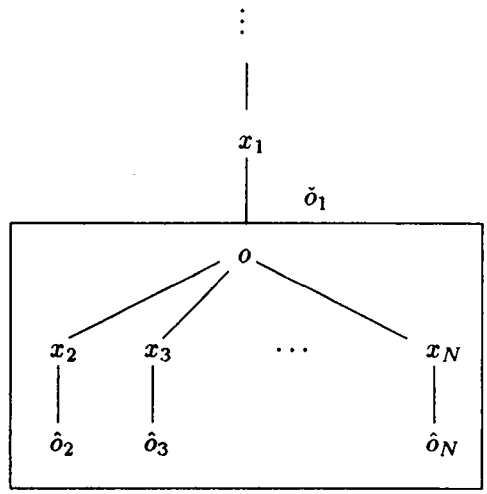

Figure 11: Merge operation.

B Smoothing

Both the two-filter and RTS smoothing algorithms can be extended to trees, although, as we now show, the structure of the tree leads to some differences in the structure of these algorithms. To begin, we consider the two-filter algorithm.

In particular, suppose that we wish to compute the smoothed estimate at a particular node x on an arbitrary tree. If we remove this x node and the arcs connected to it, we break the tree into disjoint components. The observations contained into each component can be processed recursively through the use of merging steps, so as to get a measurement summarizing the information about x contained in each subtree. All the subtree measurements can then be combined, thus yielding a smoothed observation characterizing the information about x contained in the whole tree.

In detail, consider an x-node of a tree $\mathcal{T}$, which is connected to nodes o_j with $1 \leq j \leq M$, as shown in Figure 13. The tree $\mathcal{T} - \{x\}$ obtained by removing x and the arcs connected to it from $\mathcal{T}$ is partitioned into M disconnected subtrees $\mathcal{T}_j$, where a node belongs to the subtree $\mathcal{T}_j$ if the unique path connecting this node to x goes through the node o_j. Then, let O_j be the observation obtained by o-aggregation of all observations contained in the subtree $\mathcal{T}_j$, including o_j, and let $\hat{o}_j$ be the observation obtained by extracting x from O_j and reducing the resulting observation, i.e.

$$\hat{o}_j = R\{X_x\{O_j\}\}\,. \tag{95}$$

Let also $\hat{o}^s$ be the smoothed observation obtained by extracting x from the o-aggregation of all observations in the tree, and reducing the resulting observation, so that

$$\hat{o}^s = R\{X_x\{O\}\}\,. \tag{96}$$

Then, the analog of the two-filter formula for arbitrary trees is given by

$$\hat{o}^s = R\{\oplus_{j=1}^{M} \hat{o}_j\}\,, \tag{97}$$

which shows that $\hat{o}^s$ can be obtained by combining the information about x contained in each subtree, and reducing the resulting observation.

In (97) the observations $\hat{o}_j$ can be constructed recursively by using merge operations to progressively collapse the tree $\mathcal{T}_j$ from its extremities towards node x. This can be performed systematically by giving a root structure to the tree $\mathcal{T}$, where x is selected as the root, the observations o_j are located on the first level of the tree, to the o_j's are put on the second level, etc. Then, by using merge operations to move from the higher to lower levels of the tree, one can progressively compress the observations contained in each subtree $\mathcal{T}_j$, so as to place ourselves in the situation corresponding to Figure 13.

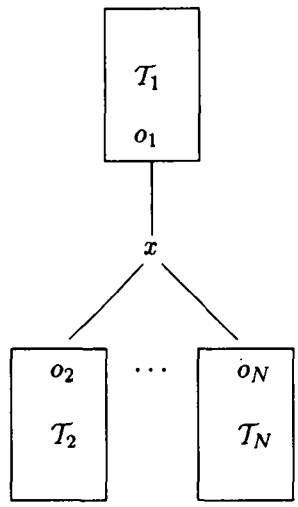

Figure 12: Smoother geometry.

Note that the generalization of the two-filter algorithm has a very simple structure with considerable symmetry. However, in contrast to the case of a linear tree, the computational structure of the algorithm is far more complex in the case of an arbitrary tree. Specifically, suppose that we wish to obtain the best estimate of x_i at every node of the tree. As we have just seen, the computation for each such node corresponds to breaking the tree into disjoint subtrees by removing the node x_i and by performing recursive processing toward x_i in each subtree. For a linear tree, this leads to *two* recursive filters, one from each extremity to the other, providing *all* of the subtree estimates required for optimal smoothing at *all* x-nodes. For an arbitrary tree, with several extremities, this is obviously *not* the case, implying that there are many more recursive filters (essentially from each extremity toward every other extremity). Note that there is considerable parallelism in this estimation structure and thus the computational burden is not prohibitive. Indeed in [27] a related algorithmic structure is developed and shown to be computationally efficient for the problem of whitening of data on such a tree in order to calculate likelihood functions. Furthermore, there are obvious simplifications if estimates are desired only at a subset of x-nodes. Nevertheless the nature of this computational structure does provide motivation for the consideration of alternatives.

One such structure is the RTS algorithm, which retains a key property that it shared

with its counterpart for linear trees but which the two-filter generalization loses for arbitrary trees. In particular, the RTS algorithm on an arbitrary tree involves recursive processing that passes through each x and o-node *exactly* twice, on a "forward" and on a "backward" sweep. To achieve this, however, we must adopt an asymmetric view of the tree by choosing one particular extremity of the tree as the "top" of the tree. If we then imagine "hanging" the tree from this one extremity – so that all other extremities are at the "bottom" of the tree, then we can define a particular forward filtering algorithm in which measurements are incorporated and merged as we move from the bottom of the tree toward the top. At each x-node, then, this forward filter combines all information in the subtree below it exactly as in (92)-(94) and Figure 12. Once we have reached the top of the tree, of course, we have extracted the reduced information for x at this root node given *all* of the data in the entire tree – i.e. we have the smoothing result at the root node. The reverse sweep of RTS then looks very much like it does in (87a) - (90) for the linear tree, except that this computation in performed recursively (and in parallel) from the root node back towards *each* bottom-level extremity. In particular, if x_{k+1} refers to a node on the tree and x_k to a node one level closer to the bottom – i.e. if x_{k+1} and x_k are connected through the observation o_{k+1} – and if x_k is closer to the bottom level, then (87a) - (90) exactly describe how the smoothed estimate at the x_{k+1} node is propagated "down" to the smoothed estimate at x_k. As we have pointed out the reader is referred to [10, 11, 12] for the details of this algorithm for the nonsingular problem on dyadic trees described in Example 5. Also examples of this algorithm for the 2-D structure of Example 4 are given in [18, 24] including a discussion in [24] of its relationship to nested dissection methods for solving sparse, locally-related linear systems of equations [25].

7 Conclusions

In this paper we have developed a general framework for deriving recursive ML estimation algorithms for problems specified by noisy linear relations describing either linear stochastic models or measurements. An xo-graph structure was associated to each estimation problem. It was then shown that if any xo-graph can be reduced to acyclic form through the use of x- or o-aggregation operations, and for any such acyclic form it is possible to derive recursive estimation algorithms for the corresponding reduced estimation problem. The recursive ML estimation algorithms we have developed rely on two elementary operations, called reduction and extraction, which can be used to compress observations, and extract the information about certain variables contained in these observations. The resulting filtering and smoothing algorithms were illustrated for both linear and arbitrary trees. These results are very general, since they apply to 1-D and multidimensional stochastic systems, systems with singular dynamics, and stochastic processes defined at multiple resolution levels. Furthermore, the procedures employed to perform reduction and extraction operations rely on numerically stable methods, of the same type as to those arising in square-root Kalman filtering, and thus yield numerically reliable estimation techniques. This general framework appears to offer considerable promise for a wide variety of estimation problems, including those for 2-D processes and those involving scale-recursive process descriptions. In addition the structures we have described for 2-D problems naturally suggest approximations that yield potentially dramatic computational savings with little loss in performance

[18, 23, 24].

References

[1] G. Chartrand and L. Lesniak, *Graphs and Digraphs*, 2nd ed. Monterey, CA: Wadsworth, 1986.

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

[3] R. P. Roesser, "A discrete state-space model for linear image processing," *IEEE Trans. Automat. Control*, vol. 20, pp. 1-10, Feb. 1975.

[4] E. Fornasini and G. Marchesini, "State-space realization theory of two-dimensional filters," *IEEE Trans. Automat. Control*, vol. 21, pp. 484-491, 1976.

[5] B. C. Levy, M. B. Adams, and A. S. Willsky, "Solution and linear estimation of 2-D nearest-neighbor models," *Proc. IEEE*, vol. 78, pp. 627–641, April 1990.

[6] R. Nikoukhah, A. S. Willsky, and B. C. Levy, "Kalman filtering and Riccati equations for descriptor systems," to appear in *IEEE Trans. Automat. Control*, vol. 37, Sept. 1992.

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

[8] P. Whittle, *Prediction and Regulation by Linear Least-Square Methods*, 2nd ed. Minneapolis, MN: University of Minnesota Press, 1983.

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

[10] K. C. Chou, A. S. Willsky, A. Benveniste, and M. Basseville, "Recursive and iterative estimation algorithms for multi-resolution stochastic processes," in *Proc. 28th IEEE Conf. Decision and Control*, Tampa, FL., Dec. 1989, pp. 1184-1189.

[11] K. C. Chou, "A stochastic modeling approach to multiscale signal processing," Ph. D. thesis, Dept. of Electrical Engineering and Computer Science, and Report LIDS-TH-2036, Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA, May 1991.

[12] M. Basseville, A. Benveniste, K. Chou, S. Golden, R. Nikoukhah, and A.S. Willsky, "Modeling and Estimation of Multiresolution Stochastic Processes," *IEEE Trans. on Inf. Theory*, Vol. 38, April 1992, pp.766-784.

[13] R. Nikoukhah, A. S. Willsky, and B. C. Levy, "Boundary-value descriptor systems: well-posedness, reachability and observability," *Internat. J. Control*, vol. 46, pp. 1715-1737, 1987.

[14] R. Nikoukhah, "A deterministic and stochastic theory for two-point boundary-value descriptor systems," Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, and Report LIDS-TH-1820, Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA, 1988.

[15] L. Chisci and G. Zappa, "Square-root Kalman filtering of descriptor systems," preprint, 1992.

[16] L. Dai, "Filtering and LQG problems for discrete-time stochastic singular systems," *IEEE Trans. Automat. Control*, vol. 34, pp. 1105-1108, Oct. 1989.

[17] X-M. Wang and P. Bernhard, "Filtrage et lissage des systèmes implicites discrets," Technical report no. 1083, Institut National de Recherche en Informatique et Automatique, Rocquencourt, France, Aug. 1989.

[18] D. Taylor, Parallel estimation on one and two dimensional systems," Ph.D thesis, Dept. of Electrical Engineering and Computer Science, and Report LIDS-TH-2092, Laboratory for Information and Decision Systems, M.I.T., Cambridge, MA, 1992.

[19] G. H. Golub and C. H. Van Loan, *Matrix Computations*, 2nd edition. Baltimore, MD: The Johns Hopkins Univ. Press.

[20] D. Q. Mayne, "A solution of the smoothing problem for linear dynamic systems," *Automatica*, vol. 4, pp. 72-92, Nov. 1966.

[21] D. C. Fraser, "A new technique for the optimal smoothing of data," D.Sc. Dissertation, Dept. Aeronautics and Astronautics, M.I.T., Cambridge, MA, Jan. 1967.

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

[23] M. Luettgen, W.C. Karl, and A.S. Willsky, "Multiscale Representations of Markov Random Fields," *IEEE Trans. on Signal Processing*, special issue on wavelets, to appear.

[24] M.M. Daniel and A.S. Willsky, "Parallel Algorithms for 2-D Noncausal IIR Filters," submitted to the 1993 IEEE Workshop on Image and Multidimensional Signal Processing.

[25] A.George and J.W. Liu, *Computer Solution of Large and Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.

[26] S.L.Campbell and C.D.Meyer, *Generalized Inverses of Linear Transformations*, Pitman, London, 1979.

[27] M.Luettgen, A.S. Willsky, and W.C. Karl, " Likelihood Calculation for a Class of Multiscale Stochastic Models," in preparation.

Unité de Recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

Unité de Recherche INRIA Lorraine Technopôle de Nancy-Brabois - Campus Scientifique
615, rue du Jardin Botanique - B.P. 101 - 54602 VILLERS LES NANCY Cedex (France)
Unité de Recherche INRIA Rennes IRISA, Campus Universitaire de Beaulieu 35042 RENNES Cedex (France)
Unité de Recherche INRIA Rhône-Alpes 46, avenue Félix Viallet - 38031 GRENOBLE Cedex (France)
Unité de Recherche INRIA Sophia Antipolis 2004, route des Lucioles - B.P. 93 - 06902 SOPHIA ANTIPOLIS Cedex (France)

EDITEUR
INRIA - Domaine de Voluceau - Rocquencourt - B.P. 105 - 78153 LE CHESNAY Cedex (France)

ISSN 0249 - 6399