Distributional Kalman Filters for Bayesian Forecasting and closed form recurrences

J.Q.Smith and G. Freeman
The University of Warwick

July 4, 2010

Abstract

Over the last 50 years there has been an enormous explosion in developing full distributional analogues of the Kalman filter. In this paper we explore some of the ways analogues of the orginal second order processes discovered by Kalman have their analogues in Bayesian state space models. Many of these analogues need to be calculated using numerical methods like MCMC so they retain, or even enhance the descriptive power of the Kalman Filter, but at a cost of transparency. However, if the analogues are drawn properly, elegant recurrence relationships - like those of the Kalman Filter - can still be developed that apply at least for one step ahead forecast distribution. In this paper we explore the variety of ways such models have been built, in particular with respect to graphical time series models..

1 Introduction

Kalman's seminal paper and its many analogues spawned enormous activity over a wide range of environments. In this paper we will restrict our attention to a small subset of these fields: the impact of his work on the Bayesian analysis of discrete time series models, the relationship between Kalman's work and recent developments in graphical modeling and models of causation and control.. Even surveying all the important work in these three area would be a hopeless task in this short review. So I hope the reader will forgive us for focusing on those modeling issues about which we have a particular interest. Even then we will pay special attention to multivariate time series whose structure can be represented by graphs where it is possible to describe explicit closed form recurrences for the series and their causal extensions to controlled models. The motivating examples we use will be drawn from business, social and environmental applications of this technology.

There are several reasons why the Kalman filter has such an impact on the areas mentioned above. First, and possibly most importantly, it was one of the first to explicitly recognise the efficacy of expressing a time series in terms of an underlying process which represented what was actually driving what was seen. He explicitly separated this from the measurement of the process which was imperfect and subject to error. Yes, he only considered second order processes with linear dynamics but we will see below that these naturally extended into full distributional Gaussian state space time series: see below. These Dynamic Linear Models in turn not only catalyzed the development of hierarchical linear models which form the bedrock of a large proportion of Bayesian analyses we see today, but also directly enabled a systematic study of non-Gaussian dynamic state space models which are now so widely studied especially in the modeling of financial time series.

The second impact his work had was the recognition of the existence of simple recurrences for both the one step ahead forecasts, and state estimates. The closed form solutions made the impact of hyperparameters clear in terms of their effect on adaptive coefficients, forecast variances as well as providing a narrative for why the formulae made logical sense. In particular it became possible to appreciate why certain ad hoc exponential smoothers were so successful in practice by mapping these on to values of hyperparameters in the Kalman filter where they had an interpretation in terms of the parameters of the terms in the transition equations and the parameters of the sampling error. Within the context of Bayesian Dynamic modelling, because of the excitement about the implementability of numerical methods round his framework this useful feature has been rather neglected. However as researcher become more ambitious, modeling time series over very high dimensional spaces or selecting models over enormous model spaces the efficacy of closed form recurrences is once again being recognized, not only to facilitate the necessary fast computation needed for this type of modeling but also to ensure a harmonious interpretation of process across an otherwise very heterogeneous domain.

Rather later, after the full distribution analogues of the Kalman filter had been developed, researchers became aware that one reason the recurrences could exhibit a closed form was the existence of the conditional independence structure implicit in the Markovian assumptions of the distributional Kalman Filter. The distributional Kalman filter is a Bayesian Network and this was one of the first non-trivial graphical models to be widely studied. The recurrences of the decomposable Gaussian Bayesian Network can be seen as a rather trivial extension of backward and froward recurrences associated with the Kalman filter. So again the Kalman filter was the forerunner of another completely different but large class of models which is now extremely widely applied. Over the last twenty years or so many authors have explored dynamic versions of BN's. Developments of the Kalman filter structure have been intrinsic to the success of many of these methods.

Most recently there has been a vigorous study of causal systems. Somewhat belatedly it has been appreciated that ideas of causation are best viewed within a dynamic framework where causal hypotheses are expressed in terms of controls on the dynamic system. Again the distributional Kalman filter has been a natural starting point from which to develop a proper understanding of causal models. First its very structure was designed to separate the actual process from the observed process. Because of this it naturally admits an extension to model a controlled process where control can be enacted by manipulating the system error to take certain values. Indeed in the early applications of the technology this embedding was often immediately valid because of context. So early applications of Kalman filters were in fact often what we would call now applications of causal models.

Second causal models Pearl (2000) demand that factorisation formulae of a joint density of variables in a controlled system are a simple transformation of these factors as they appear in the uncontrolled system. Happily the Kalman filter formulae for prediction under control respect exactly this sort of transformation. The controlled Kalman filter is beginning to have a strong impact on this area of study as well.

Our story will begin with reviewing in the next section how the Kalman filter was transformed into a full Bayesian model through its development into a Dynamic Linear Model. We will then outline some developments of closed form recurrences associated with distributions other than the Gaussian and the links with exponential smoothing. We continue by exploring the use of implicit conditional independences in the Kalman filter that enabled its links with hierarchical structures to be better appreciated. This conditional independence structure let to the discovery of new classes of multivariate models which on the one hand expressed useful qualitative hypotheses and on the other could exploit the implicit factorizations induced through the state space structure to obtain useful recurrences on the system. We will conclude by illustrating how control - interpreted in a Kalman filter way - can be used to explain two sorts of causal models of processes: one based on a model of the oil market and the other for forecasting the effects of different educational programmes.

2 Dynamic Linear models

In our judgment the power of the Kalman filter to model in a business environment only became fully apparent when the recurrences were reinterpreted by Bayesians as a full probabilistic description of a Gaussian process. This was first mooted in a seminal paper by [12]. From a technical point of view this paper simply took the Kalman filter and assumed all variables in the system were Gaussian. Thus in their notation the univariate Dynamic Linear Model (DLM) was given by and observation equation

$$Y_t = \boldsymbol{F}_t(\boldsymbol{x}_t)\boldsymbol{\theta}_t + v_t \tag{1}$$

and a system equation

$$\boldsymbol{\theta}_t = G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w}_t \tag{2}$$

where the observational and system errors $\{v_t, \boldsymbol{w}_t : t = 1, 2, 3, \ldots\}$ independent Gaussian distributions with zero mean observation variance V_t and system error covariance matrix $\boldsymbol{W}_t$. Use the usual convention that y^t denotes the vector of observations $(y_1, y_2, \ldots, y_t)$. The row vector $\boldsymbol{F}_t(x_t)$ was a vector of known coefficients as a function of a vector of covariates $\boldsymbol{x}_t$. When these recurrence relationships are adjoined to a prior which sets $\boldsymbol{\theta}_0 \backsim N(m_0, C_0)$ then it is easily seen that conditional on the parameters of $\{\boldsymbol{F}_t(\boldsymbol{x}_t), G_t, V_t, \boldsymbol{W}_t\}$ the general recurrences could then be written as $\boldsymbol{\theta}_t|\boldsymbol{y}^t \sim N(\boldsymbol{m}_t, C_t)$ and $Y_t|\boldsymbol{y}^{t-1} \sim N(f_t, Q_t)$

$$\begin{aligned}\boldsymbol{m}_t &= G_t\boldsymbol{m}_{t-1} + \boldsymbol{A}_t\left(y_t - f_t\right)\\ C_t &= G_tC_{t-1}G_t^T + W_t - \boldsymbol{A}_tQ_t\boldsymbol{A}_t^T\end{aligned}$$

where $f_t = \boldsymbol{F}_tG_t\boldsymbol{m}_{t-1}$,and [Do we need a vector form of this!!]

$$\begin{aligned}\boldsymbol{A}_t &= [G_tC_{t-1}G_t^T + W_t]\boldsymbol{F}_t^TQ_t^{-1}\\ Q_t &= \boldsymbol{F}_t\left(C_t + A_tQ_tA_t^T\right)\boldsymbol{F}_t^T + V_t\end{aligned}$$

Although technically trivial these distributional assumptions made the Kalman filter recurrences apply to relationships between conditional means in a full continuous discrete time possibly non -stationary stochastic process. This was a philosophical leap which led to many ramifications and which are still being worked through 40 years on.

The first and perhaps the most important use of this augmentation was a practical one. Until this point, for largely technical reasons - most statistical methodology had been performed on long time series which were assumed to either be stationary or whose first or second differences were stationary. At least within a business domain these classes of models were of a rather limited use. The majority of interesting time series were being continually interrupted by exogenous shocks which often not only disturbed the development but also the generating process of the series in fundamental ways. There was often domain knowledge to explain these disruptions and drive hypotheses about their consequences, but there was no seamless way in which the usual time series technology could be combined with this knowledge to produce formal adaptations of the models. Furthermore current simple business time series models based on ideas associated with exponential smoothing and were so fundamentally non-stationary in their description did not interface well with Box-Jenkins model formulations where non stationarity could only be accommodated indirectly through differencing.

However the distributional analogues of the Kalman recurrences were ideally suited to address business series of this type. To begin with the specification of a process through a recurrence meant that disruptive events and controlled interventions could easily be dealt with within this formalism. A shock would change the state mean vector, possibly its variance and also possibly the data generation after the event through a modification of $\{\boldsymbol{F}_t(\boldsymbol{x}_t), G_t, V_t, \boldsymbol{W}_t\}$ after the time of the shock. Furthermore advances in subjective probabilistic Bayesian inference gave a *formal* framework within which this expert judgement could be accommodated into the system. And note in particular that there is nothing in the DLM framework which gives special status to stationarity. The outworkings of these advantages have continued to be vigorously exploited in an enormous number of different domains by many different authors. More formally these ideas are finally finding their application in the description of causal models: a current research domain for many working on the interface of artificial intelligence Bayesian inference and social and business modeling. We wills ee examples of these causal models later in the paper.

The links with Bayesian inference were profound partly because the new formulation emphasized the forecasting rather than parameter estimation. Of course problems of hyperparameter estimation within the components$\{\boldsymbol{F}_t(\boldsymbol{x}_t), G_t, V_t, \boldsymbol{W}_t\}$ is just as hard or harder than their Box - Jenkins analogues. and until the advent of numerical methods was a significant challenge even in the simplest processes. Indeed the fact that some of these parameters are unidentifiable or almost unidentifiable continues to present interesting challenges. However even here there are some advantages in the Kalman filter parametrisation. First in a very large class of structural models [Harvey][29], both $\boldsymbol{F}_t(\boldsymbol{x}_t), G_t$ can be treated as known, the latter often being a matrix all of whose entries are either 0 or 1. Second for a univariate series and fixed $\boldsymbol{F}_t(\boldsymbol{x}_t), G_t, \boldsymbol{W}_t V_t^{-1}$ - the last being a noise to signal ratio, the observation variance V_t can be estimated using an inverse gamma conjugate prior and the series can then still be updated in closed form, the one step ahead forecast distributions now having a student t distribution. Finally there are expedient choices of the matrix $\boldsymbol{W}_t V_t^{-1}$ which chooses these from a one dimensional subclass parametrized by a single parameter (or a small number) of discount factors which in special cases correspond to proven exponential smoother models [29]. The process can then be mixed over these parameters - see below - or be estimated by very fast numerical techniques. Alternatively we can simply estimate using fast MCMC or particle filter methods, whenever more structured covariance structures are posited. So univariate Gaussian series are relatively straightforward to implement and there is a great deal of code now available to help the user to implement these methods. An excellent review of more recent developments stemming from this stream of research can be found in [29] and other work on distributional Kalman filters from a somewhat different perspective in [7]

One final advantage of this type of Bayesian use of the Kalman filter is that the forecasting issues were much more easily addressed and provided a degree of flexibility to address the problem at hand. In a business setting a forecasting model often required the selection of an appropriate act. Within the Bayesian formalism this act simple demanded that the expectation of a suitable utility function should be maximized. So estimation could be customised to a given purpose immediately within this framework..

3 Closed form recurrences for Non Gaussian analogues of Kalman filters

Although the assumption that the distribution of all variables in the system were Gaussian led to an elegant analysis it was also recognised by Harrison and Stevens [12] that there were many scenarios where such an assumption was untenable even for univariate series. A simple extension to a wider class of models which still retained closure was the class of mixture models called Multiprocess models (Class 1). These simply assumed that one of a fixed number of Gaussian distriibutions governed the process: it was just that the modeler was uncertain which one was correct. Working within the Bayesian framework the modeler then simply assign a prior probability to each model and made predictions using the corresponding mixture of Gaussian densities. And of course both the component Gaussian distributions and their probabilities could be calculated in closed form. Interestingly many particle filter methods of time series analysis are based on this simple idea, albeit with considerable extra finesse. Class 2 Multiprocess models were also introduced at this time which gave a sequential approximate method which even allowed for models $\{\boldsymbol{F}_t(\boldsymbol{x}_t), G_t, V_t, \boldsymbol{W}_t\}$ to change within a restricted class but at *any* time. There is now an excellent recent review of these mixture methods given in [10].

Of course modeling with mixtures of Gaussian distributions is not universally applicable. However in settings which were not time series there were interesting closed form recurrences for posterior distributions and their one step ahead predictives which we might have hoped would translate into recurrences on time series for $\boldsymbol{\theta}_t|y^t$ and $Y_t|y^{t-1}$. However within the usual formulation this is usually impossible. This is because, except in a few exceptional circumstances - Bather identified these very early on - the adding on of independent errors in the system equation destroyed most natural conjugacy. So in most cases if we insist on using this type of additive formulation of the distributional filter then we have to fall back on to numerically intense and approximate methods to study this class.

However some reflection suggests that whilst in engineering and physical applications the system error has a clear meaning, in for example a business environment where a state θ might represent something like the true level of desirability of a product, then *adding an error* may not be the only way - and possibly not even the natural way - of increasing uncertainty from one time period to the next. Next suppose we assume that each the distribution of each observation Y_t is independent of everything else given the vector of parameters $\boldsymbol{\theta}_t$, thus formally that $Y_t \amalg \boldsymbol{Y}^{t-1}, \boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}_t$, $t = 1, 2, \ldots, n$ Notice that this is a property of the original Gaussian Kalman filter above. Then the only feature we will ever be able to observe directly and is the observed series unambiguous the joint mass function or density $p(\boldsymbol{y}^T)$ of the observations $\boldsymbol{Y}^T$ up to any future time T , since

$$p(\boldsymbol{y}^T) = \prod_{t=2}^{T} p(y_t|\boldsymbol{y}^{t-1})p(y_1)$$

where

$$
\begin{aligned}
p(y_1, \boldsymbol{\theta}_1) &= p(y_1|\boldsymbol{\theta}_1)p(\boldsymbol{\theta}_1) \\
p(y_t, \boldsymbol{\theta}_t|\boldsymbol{y}^{t-1}) &= p(y_t|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|\boldsymbol{y}^{t-1})
\end{aligned} \tag{3}
$$

so that in particular from marginalisation we have what we need, viz

$$
\begin{aligned}
p(y_1) &= \int p(y_1, \boldsymbol{\theta}_1) d\boldsymbol{\theta}_1 \\
p(y_t|\boldsymbol{y}^{t-1}) &= \int p(y_t, \boldsymbol{\theta}_t|\boldsymbol{y}^{t-1}) d\boldsymbol{\theta}_t
\end{aligned}
$$

Some authors e.g. [3] have even demanded that a Bayesian model is solely defined by what it can predict $p(\boldsymbol{y}^T)$. Certainly the only parameters that are unidentifiable are ones which appear in this joint density: everything else about the model cannot be learned about from the data. Since any model will give us $\{p(y_t|\boldsymbol{\theta}_t), t = 1, 2, \ldots\}$ and $p(\boldsymbol{\theta}_t|\boldsymbol{y}^t)$ can be obtained by conditioning out y_t from $p(y_t, \boldsymbol{\theta}_t|\boldsymbol{y}^{t-1})$ it follows that to develop a recurrence that fully specifies $p(\boldsymbol{y}^T)$ the system equation only needs to define how we obtain $p(\boldsymbol{\theta}_t|\boldsymbol{y}^{t-1})$ from $p(\boldsymbol{\theta}_{t-1}|\boldsymbol{y}^{t-1})$ $t = 1, 2, \ldots$. Of course the linear system equation 2 does this, or indeed any such a system equation whose states are any invertible function of $\boldsymbol{\theta}$. But this type of description also specifies a lot of other information about the sample paths of the states which are unobservable conditional on the densities given above. It will also typically lose conjugacy so the convenient closed form of the recurrences is lost. Are there classes of process for which this can be specified directly where the closed form of recurrences is retained?

The answer to this question is affirmative for a useful class of problem and have been well developed now for process which drift or have a *steady* evolution [29]. The *power steady models* [24], [17], .,[13] use the idea of increasing the temperature of a joint density at each time step by specifying that

$$
p(\boldsymbol{\theta}_t|\boldsymbol{y}^{t-1}) \propto \left\{p(\boldsymbol{\theta}_{t-1}|\boldsymbol{y}^{t-1})\right\}^k \tag{4}
$$

for some $0 < k \leq 1$ where the proportionality constant is uniquely determined because $\int p(\boldsymbol{\theta}_t|\boldsymbol{y}^{t-1})d\boldsymbol{\theta}_t = 1$ have proved particularly useful in a number of scenarios. This type of direct specification of the state transition conditionals does not specify the full joint distribution of states and observations. However those parts of the joint distributions of the states not determined in the evolution are not identifiable and irrelevant to the observed process.

These evolutions have a number of advantages. They give the same steady state recurrences as the conventional Steady DLM [29].but usually also admit conjugate evolutions when the analogous non-time varying problem has this property, with statistics replaced by familiar exponentially weighted moving average analogues which makes them accessible and interpretable to many users. So links with other distributional forms of demonstrably useful classes of smoothers are formalised in this way. In a multivariate setting logical constraints will be preserved with time as well as all independences and many conditional independences existing from the previous time slice, and various characterisations of this type of process exist. Finally the evolution can be characterised [**?**], [26]. has an invariant decision based, or using linear shrinkages of either Kullback Leibler distances or local DeRobertis distances see Freeman and Smith discussed below. Of course they are not universally appropriate and the class does not admit an obvious continuous time analogue.

These methods have recently found new applications when modeling very high dimensional non-Gaussian time series because its closed form makes fast computation, needed for model selection and estimation feasible see for example [**?**], [22], [2] and [23] an example outlined below.

4 Bayesian Networks and Space Dynamic Graphical Models

It has long been recognized that the structure of a distributional Kalman filter over n time periods, whether its observations are univariate or multivariate can be expressed as a Bayesian Network defined on a sequence of vectors $\boldsymbol{Y}^T \triangleq (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T)$ and $\boldsymbol{\theta}^T \triangleq (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_T)$. Thus it is straightforward to check that the conditional independences explicit in the general distributional Kalman filter can be written in the form

$$\begin{aligned} \boldsymbol{Y}_t \amalg \boldsymbol{Y}^{t-1}, \boldsymbol{\theta}^{t-1} | \boldsymbol{\theta}_t, \, t &= 1, 2, \ldots, n \\ \boldsymbol{\theta}_t \amalg \boldsymbol{Y}^{t-1}, \boldsymbol{\theta}^{t-2} | \boldsymbol{\theta}_{t-1}, \, t &= 2, 3, \ldots, n \end{aligned}$$

These are exactly the conditional independences of a valid BN whose DAG is

$$\begin{tikzcd}
\boldsymbol{Y}_1 & \boldsymbol{Y}_2 & & \boldsymbol{Y}_{t-1} & \boldsymbol{Y}_t & \boldsymbol{Y}_{t+1} & & \boldsymbol{Y}_n \\
\boldsymbol{\theta}_1 \arrow[u] \arrow[r] & \boldsymbol{\theta}_2 \arrow[u] \arrow[r] & \cdots \arrow[r] & \boldsymbol{\theta}_{t-1} \arrow[u] \arrow[r] & \boldsymbol{\theta}_t \arrow[u] \arrow[r] & \boldsymbol{\theta}_{t+1} \arrow[u] \arrow[r] & \cdots \arrow[r] & \boldsymbol{\theta}_n \arrow[u]
\end{tikzcd}$$

Various useful non-distributional results can be proved about any process whose distributions respect this dependence structure simply by evoking the d separation theorem [18],[16] that are much more obscure when proved in other ways. For example writing $\widehat{\boldsymbol{\theta}}_t = (\boldsymbol{\theta}_1, \ldots \boldsymbol{\theta}_{t-2}, \boldsymbol{\theta}_{t+2}, \ldots \boldsymbol{\theta}_n)$, $\widehat{\mathbf{Y}}_t = (\mathbf{Y}_1, \ldots \mathbf{Y}_{t-2}, \mathbf{Y}_{t+2}, \ldots \mathbf{Y}_n)$ it can immediately be proved that

$$\boldsymbol{\theta}_t \amalg \left(\widehat{\boldsymbol{\theta}}_t, \widehat{\mathbf{Y}}_t \right) | (\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t+1})$$

which in turn gives a framework for understanding retrospective analyses of parameters using non-Gaussian forms of standard DLM's.

Conditional independence models have a close relationship to the specification of model classes through the factorisations exhibited in their joint densities - see e.g. [4] - and in turn the distributional forms of system and observation equation link directly to factorisation formulae. Through this connection it can be seen that Gaussian Kalman filters with their fast algorithms for calculating were the forerunners of the fast propagation algorithms on decomposable

Gaussian Bayesian networks using Junction trees [2], [14]. Thought of in this way the states of a Kalman filter correspond to the separators on the junction tree of the Bayesian Network given above. Interesting there have also been developments of second order recurrences over trees see for example Lauritzen and [11] which exactly correspond to extensions of Kalman's original work.

This recognition of the link between state space time series and graphical models provoked the development of dynamic Bayesian Networks [5], These can be seen as a distributional Kalman filter with added conditional independences over the state transitions and time constant parameters$\{\boldsymbol{F}_t(\boldsymbol{x}_t), G_t, V_t, \boldsymbol{W}_t\}$. In the usual Gaussian form of this model the conditional independences can simply be expressed in terms of collections of zeros in the transition matrix G and the state error covariance matrix W. One highly popular class of models is the 2 time slice dynamic model 2TDM ,[6], [15]. A very simple example of this class with bivariate states $\boldsymbol{\theta}_t = (\theta_{1,t}, \theta_{2,t})$, $t = 1, 2, \ldots, n$ and univariate observations given below. This embodies the conditional independences

$$\begin{aligned} Y_t \amalg \boldsymbol{Y}^{t-1}, \boldsymbol{\theta}^{t-1}, \theta_{2,t} | \theta_{1,t},\, t &= 1, 2, \ldots, n \\ \theta_{1,t} \amalg \boldsymbol{Y}^{t-1}, \boldsymbol{\theta}^{t-2} | \boldsymbol{\theta}_{t-1},\, t &= 2, 3, \ldots, n \\ \theta_{2,t} \amalg \boldsymbol{Y}^{t-1}, \boldsymbol{\theta}^{t-2}, \theta_{1,t-1} | \theta_{2,t-1},\, t &= 2, 3, \ldots, n \end{aligned}$$

and is depicted below, or more conventionally simply by the third fourth and fifth column of this BN.

$$\begin{tikzcd}
 & Y_{t-1} & Y_t & & Y_n \\
\cdots \arrow[r] & \theta_{1,t-1} \arrow[u] \arrow[r] & \theta_{1,t} \arrow[u] \arrow[r] & \cdots \arrow[r] & \theta_{1,n} \arrow[u] \\
\cdots \arrow[r] \arrow[ur] & \theta_{2,t-1} \arrow[r] \arrow[ur] & \theta_{2,t} \arrow[r] \arrow[ur] & \cdots \arrow[r] \arrow[ur] & \theta_{2,n}
\end{tikzcd}$$

At least in its linear form here we could simply specify this as a distribution Kalman filter with

$$G = \begin{pmatrix} g_{11} & g_{12} \\ 0 & g_{22} \end{pmatrix}, W = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}$$

There are many examples of the use of somewhat more complicated two time slice models than the one illustrated above (see e.g. [15]).Here the challenge is to learn the values of the hyperparameters. Even in the simple example above the likelihood over the hyperparameters $(g_{11}, g_{12}, g_{22}, W_1, W_2, V)$ is known to be an unpleasant function and tends to be rather flat except when sample sizes are large. So even in this simple example estimation is hard. The technicians in this area tend to use an approximate sequential algorithm here, because all formal closure is lost within this class.

The necessity for estimating hyperparameters $\{\boldsymbol{F}_t(\boldsymbol{x}_t), G_t, V_t, \boldsymbol{W}_t\}$ tends to cause a lack of closure. We have discussed earlier some partial solutions to these problems when the series is univariate. But problems are particularly acute problem once observations are multivariate where all but the most symmetric model suffer from this deficiency so to implement these models often very sensitive numerical techniques and sequential approximations need to be applied.

However a partial solution to these problems using as stochastic version of graphical modeling techniques to piece together many univariate Gaussian Kalman filters - explicitly regression dynamic linear models, so that - at least conditional on a few discount parameters governing noise to signal ratios of the component models - recurrences of these multivariate models are all exact. These are called Multiregression Dynamic Model (MDM), [20], [21]. This has a different conditional independence structure where the probabilities or regression parameters are linked directly to parents of each component in the time series. The simple typical MDM on a 4 - vector time series $\boldsymbol{Y}_t = (Y_t(1), Y_t(2), Y_t(3), Y_t(4))$, $t = 1, 2, \ldots$ is given below

$$\begin{tikzcd} Y_t(1) \arrow[r] & Y_t(3) \arrow[d] \\ Y_t(2) \arrow[r] \arrow[ur] & Y_t(4) \end{tikzcd}$$

Here the graph means the following: $\{Y_t(1)\}_{t\geq 1}$ and $\{Y_t(2)\}_{t\geq 1}$ are assumed to be mutually independent time series governed by a DLM $\{Y_t(3)\}_{t\geq 1}$ is also a DLM $\{\boldsymbol{F}_t(\boldsymbol{y}^t(1), \boldsymbol{y}^t(2)), G_t, V_t, \boldsymbol{W}_t\}$ but where $\boldsymbol{F}_t(\boldsymbol{y}^t(1), \boldsymbol{y}^t(2))$ is alone is a function of the observations of the two series$\{Y_s(1)\}_{s\leq t}$ and $\{Y_s(2)\}_{s\leq t}$ up to and including observations at time t. Finally $\{Y_t(4)\}_{t\geq 1}$ is also a DLM $\{\boldsymbol{F}_t(\boldsymbol{y}^t(2), \boldsymbol{y}^t(3)), G_t, V_t, \boldsymbol{W}_t\}$ but where $\boldsymbol{F}_t(\boldsymbol{y}^t(2), \boldsymbol{y}^t(3))$ is a function of the observations of $\{Y_s(1)\}_{s\leq t}$ and $\{Y_s(2)\}_{s\leq t}$ but not $\{Y_t(1)\}_{t\geq 1}$.

What is a surprising and useful property of these processes is that if the state vectors $\{\boldsymbol{\theta}_0(1), \boldsymbol{\theta}_0(2), \boldsymbol{\theta}_0(3), \boldsymbol{\theta}_0(4)\}$of each of the 4 series are assume to be mutually independent at time 0 then $\{\boldsymbol{\theta}_t(1)|\boldsymbol{y}^t, \boldsymbol{\theta}_t(2)|\boldsymbol{y}^t, \boldsymbol{\theta}_t(3)|\boldsymbol{y}^t, \boldsymbol{\theta}_t(4)|\boldsymbol{y}^t\}$will also be independent for all $t = 1, 2, \ldots$. This in turn means that the one step ahead Kalman Filter recurrences for each of the component series after observing .and process are closed and it is these vectors of parameters which are each given their own independent dynamic process are also valid. This gives a rich class of dynamic graphical models, whose graphs can be naturally interpreted in terms of dependence structures and whose time series structure is very easy to analyse: see for example Queen (2009) for a simple example of these, and series B paper for a continuous analogue. The class of models is particularly suitable for modelling series with a fast dynamic over the components: for example when a high value of $Y_t(1)$ and $Y_t(2)$ has a tendency trigger almost instantaneously a high value in $Y_t(3)$ which in turn tends to trigger a high value of $Y_t(4)$ in the example above.

Because of the Kalman recurrences, in the example above conditional on the hyperparameters, the conditional predictive densities $\left\{y_t(1)|\boldsymbol{y}^{t-1}, y_t(2)|\boldsymbol{y}^{t-1}, y_t(3)|\left(\boldsymbol{y}^{t-1}, y_t(1), y_t(2)\right), y_t(4)|\left(\boldsymbol{y}\right.\right.$ are each Gaussian. However the joint distribution of $\boldsymbol{y}_t|\boldsymbol{y}^{t-1}$ certainly is not which explains why this class of processes does not suffer the same problem as the usual multivariate analogues of the distributional Kalman filter. However if the functions $\boldsymbol{F}_t(\boldsymbol{y}^t(1), \boldsymbol{y}^t(2)), \boldsymbol{F}_t(\boldsymbol{y}^t(2), \boldsymbol{y}^t(3))$ are functions of these series only through the values of $(y_t(1), y_t(2))$ and $(y_t(2), y_t(3))$ and are polynomial, for example

$$\begin{aligned}\boldsymbol{F}_t(\boldsymbol{y}^t(1), \boldsymbol{y}^t(2)) &= (1, y_t(1)y_t(2)) \\ \boldsymbol{F}_t(\boldsymbol{y}^t(2), \boldsymbol{y}^t(3)) &= (1, y_t(2), y_t(3))\end{aligned}$$

respectively then all the joint moments of this one step ahead predictive to any order are trivial to calculate using standard results in probability theory [21]. Of course, because of the potential complexity of these joint predictive densities, retrospective analyses are extremely complex to perform accurately. But for one step ahead forecasting these processes are simple to analyse.

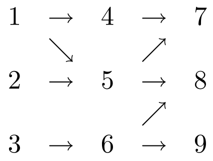
5 Causal models and controlled Kalman Filters

Over the last fifteen years or so, there have been enormous activity on the study of causation using probabilistic graphical models. However much of the recent work, often associated with medicine and science, have tended to focus attention on cross sectional models where causation acts over a single time point. More recently interest has refocused back on to the proper domain of expression of causal relationships: the area of multivariate time series.

Many authors and particularly Pearl[19] and Dawid see e.g. have come to realise that one of the most effective ways of modeling causation is through its relation to controls. But within the Kalman filter formulation of process this has always been the way causality has been described. Thus for example in causal models of the movement of a missile, where states represent the acceleration in different coordinates an action to change this vector by manipulating the system error to reflect the control whilst leaving the other elements of the system - describing the environment and measurement errors - intact.

This expressiveness of extending an the description of an uncontrolled time series to one which was manipulated was recognized [12] very early in the study of distributional forms of Kalman filters Furthermore by West and Harrison (1982) had developed these methods formally so they could be applied to a wide class of business time series models which were subjected to many different types of control or subjected to many different types of external shocks.

Within certain markets where supply is demand led and passes through certain agents causal structures are a little more subtle. However here by appropriately defining states, it is still possible to use the structure of Kalman filters to model causal relationships. Thus suppose we have a commodity can pass along the routes in the graph given below where $\{1, 2, 3\}$ are drillers, $\{4, 5, 6\}$ are refiners and $\{7, 8, 9\}$ distributors and suppose for simplicity that it takes exactly 1 time period to transport the commodity from a node to one leading from it.



The typical information we have available at any one time is the total amount of the commodity at some subset of the traders. The trick here is to label as states the source to sink paths that a commodity might pass along. Here the index at each start time t of these would be $\{147, 157, 158, 257, 258, 368, 369\}$. If we could observe take a snap shot of the commodity at a given time t - with error we would take an observation then we obtain a Kalman filter but - because of the transport delays - with some of the information hidden for up to 2 time periods. The recurrences of the standard Kalman filter are then slightly adapted but nevertheless stay in closed form. In a demand led market the economy will require the sums of the commodity to change stochastically through refiners

By using this type of parametrisation The effect of a causal change here, like the closing of a well, the withdrawal of a refiner can now be modelled in terms of its effects on the flows from well to distributor. Thus if well 1 were to close then the flows indexed by $147, 157, 158$ would all be set to zero. and these amounts would then need to be compensated by increased flows from the other two wells. so that the continued demands will be satisfied. The appropriate choice of these models will depend on context but are discussed in [27] and L's thesis. The point is here that the flexibility of the Kalman filter where states can be variously defined allows for causal effects where there are systematically delays in the reaction of the process: something that is not possible using usual graphical machinery. So this represents yet another albeit indirect application of Kalman's technology.

Our final example briefly previews a recent development on modelling high dimensional discrete time series on trees. Although this class of models of a very complex and structured domain lie at the end of a chain of development of ideas discussed here, it uses all the ideas so far attributed as having their root in Kalman's work. These are:

1. the creative supplementation of an observed process with states whose evolution can be well defined.

2. the exploitation of conditional independences between states to obtain closed form conjugate updating.

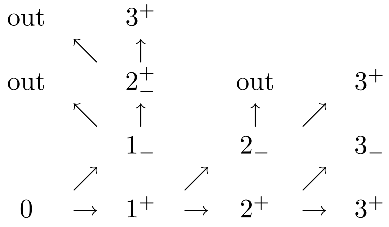
3. the use of power steady forms of evolution and Multiprocess modelling to generalise these the give closed form Kalman like recursions.

4. the exploitation of the explanatory power of the Kalman filter to extend its domain so that it applies equally well to domain where there is control or disruption.

To illustrate this type of model consider the following problem where students take a sequence of yearly courses where pass rates are supposed to be comparable but where students are thrown out of the programme if they fail two consecutive exams, and that each student attends one course a year. Interest is on the development of the programme over the years, the differences in the success rates in the different courses on offer and to forecast the effects of changes of teaching programmes and syllabi might cause. The tree below represents a very simple three year programme like this where vertices with a + superscript represent a student passing their last course. In any year we observe the numbers who pass or fail each year for each of the 6 categories of student labelled by the interior vertices of the tree.

$$\{0, 1_-, 1^+, 2^+_-, 2_-, 2^+\}$$

which are respectively, the starting students, those who have failed the first year, those who have passed Year 1, those having failed Year 1 passed Year 2, those who passed Year 1 but failed Year 2 and those who passed in their first two years. The leaves label the different possible educational records. Assume that students in a given fixed year are exchangeable within these categories. We would now like to treat the problem as a yearly time series on the vector of the probability of a student passing in each category in each given year.



The edges of the graph above can be labeled by these probabilities therefore which in Kalman filter terms will define our states. The set of possible models on the tree above might include different hypotheses that some of the 6 probabilities can be partitioned so that interior vertices in the same cluster in the partition have the same pass probabilities. The factorization of the probabilities of the tree means that all these component probabilities associated with each given partition, after complete sampling at each year, remain *independent* year on year: a typical exploitation to conditional independences induced by the Kalman state space definition we have used. Allowing a steady drift over time for each of the cluster probabilities and giving them independent beta distributions and placing a prior distribution on each possible cluster hypotheses allows us to update the probabilities of the next year's student performance as a Class 1 steady Mixture Model. This is a hybrid of two of the models discussed above and exploits Kalman filter closed form analyses. Class 2 Multiproces models can also be modified to this context to allow transitions over time from one partition to another.. Finally, if we are interested in investigating whether changing a syllabus has made a course less easy to pass than others in its current cluster, we simply need to extend this distribution Kalman filter into a controlled model where we replace the variable labeled by the interior vertex by a new controlled one: i.e. augment the distribution after control locally at its given time and state whilst retaining the distribution of the other states: just as we do in a controlled Kalman filter.

Because these procedures or all closed form analyses of large student programs with hundreds of courses can still be studied using this model class. For a detailed discussion of this type of model see Freeman and Smith (2011), Freeman and Smith (2010) and Freeman(2010).

6 Conclusions

The impact of Kalman has been and continues to be enormous. In particular we believe that his brilliant concept of separating states from observation and utilizing Markov properties, to better explain the process, better estimate the process and better embed the models into more complex domains admitting shocks and controlled interventions will have an enduring influence on the construction of forecasting models well into the future.

References

[1] Atwell and Smith(1991) "A Bayesian forecasting model for sequential bidding" J.of Forecasting, 10, 565 - 577

[2] R.G.Cowell,A.P. Dawid,S.L. Lauritzen and D.J.Spiegelhalter (1999) "Probabilistic Networks and Expert Systems" Springer

[3] Dawid, A.P. (1992) "Prequential analysis, stochastic complexity and Bayesian inference" Bayesian Statistics 4 (Eds Bernardo et al) Oxford University Press 109 -125

[4] Dawid, A.P., Studený, M., 1999. "Conditional products: an alternative approach to conditional independence". In Heckerman, D., Whittaker, J. (Eds.), Artificial Intelligence and Statistics 99 Morgan Kaufmann Publishers, S. Francisco, 32–40

[5] Dahlhaus, R. and Eichler, M. (2003), Causality and graphical models for time series. In: P. Green, N. Hjort, and S. Richardson (eds.), Highly structured stochastic systems. University Press, Oxford, pp. 115-137.

[6] Dean,T. and Kanazawa, K. (1988) Probabilistic Temporal Reasoning, Proc. AAAI-88, *AAAI*, 524-528

[7] Durbin, J. and Koopman,S. J. (2001) " Time Series Analysis by State Space Methods" Oxford University Press

[8] Eichler, M. (2006), Graphical modelling of dynamic relationships in multivariate time series. In: M. Winterhalder, B. Schelter, J. Timmer (eds), Handbook of Time Series Analysis, Wiley-VCH, Berlin, pp. 335-372.

[9] Eichler, M. (2007), Granger-causality and path diagrams for multivariate time series. Journal of Econometrics 137, 334-353

[10] Fruthwirth-Schnatter, S. (2006) "Finite Mixture and Markov Switching Models" , Springer Verlag, New York,

[11] Goldstein, M. and Wooff, D. (2007) "Bayesian Linear Statistic: Theory and Methods" Wiley

[12] Harrison, P.J. and Stevens, C.F. (1976) "Bayesian forecasing"(with discussion) J.R.Statist .Soc B, 38, 205 -247

[13] Ibrahim, J.G. and Chen, M.H.(2000) "Power prior distributions for regression models" Statistical Science, 15, 46 -60

[14] Jensen F.V. and Nielsen, T.D. (2007) "Bayesian Networks and Decision Graphs"(2nd edition) Springer Verlag, New York

[15] Koeller, D. and Lerner, U. (1999) Sampling in Factored Dynamic Systems in "*Sequential Monte Carlo Methods in Practice*"

[16] Lauritzen S.L. (1996) "Graphical models". Oxford Science Press, Oxford, 1^{st} edition.

[17] Peterka, V. (!981) "Bayesian system identification". In: Trends and Progress in System Identification, P. Eykhoff, Ed., p. 239-304. Pergamon Press, Oxford

[18] Pearl,J. (1988) Probabilistic Reasoning in Intelligent Systems San Mateo: Morgan Kauffman

[19] Pearl,J. (2000). Causality. models, reasoning and inference. Cambridge University Press, Cambridge.

[20] Queen, C.M. and Smith, J.Q. (1992). "Symmetric Dynamic Graphical Chain Models". Bayesian Statistics 4. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (Eds.). Oxford University Press, 741-751.

[21] Queen, C.M., and Smith, J.Q. (1993). "Multi-regression dynamic models". J.R. Statist. Soc. B, Vol.55, No.4, 849-870.

[22] Queen, C.M., Smith, J.Q. & James, D.M. (1994). "Bayesian Forecasts in markets with overlapping structures". Int. J. of Forecasting, 10, 209-233.

[23] Rigat, F. and Smith, J.Q. (2009) "Non-parametric dynamic time series modelling with applications to detecting neural dynamics" The Annals of Applied Statistics (to appear)

[24] Smith, J.Q.(1979) "A generalisation of the Bayesian steady forecasting model" J.R.Statist. Soc . B 41, 375 -87

[25] Smith, J.Q. (1981)."Search Effort and the Detection of Faults" B.J. of Mah. and Stat. Psy, 34, 34, 181 -193

[26] Smith, J.Q. (1992). "A comparison of the characteristics of some Bayesian forecasting models". International Statistical Reviews, 60,1, 75-87.

[27] Smith, J.Q. and Figueroa-Quiroz, L.J. (2007) "A Causal Algebra for Dynamic Flow Networks" in "Advances in Probabilistic Graphical Models" Eds P. Lucas, J.A.Gamez, and A. Salmeron, Springer, 39 -54

[28] P. Spirtes, C. Glymour and R. Scheines (1993). Causation, Prediction, and Search. Springer-Verlag, New York.

[29] West, M. and Harrison, P.J.(1997) "Bayesian Forecasting and Dynamic Models" Springer.