IDIAP RESEARCH REPORT

The Auxiliary Variable Trick for deriving Kalman Smoothers

David Barber [a]

IDIAP–RR 04-87

Decemeber2004

submitted for publication

[a] IDIAP Research Institute, Martigny, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-87

The Auxiliary Variable Trick for deriving Kalman Smoothers

David Barber

Decemeber2004

submitted for publication

Abstract. We present a Forward-Backward Kalman smoother derivation that functions for small observation noise. Whilst this smoother can be found by judicious transformation of the standard Forward-Backward equations, we introduce an auxiliary variable trick which greatly simplifies the derivation, based on the probabilistic interpretation of the Kalman Filter, allowing one to work directly with moments of the distribution. The trick is of potential interest for the simple derivation of smoothing type inference in other related systems.

1 Introduction

The Kalman Filter[1, 2], a discrete-time linear Gaussian state space model, may be written as a coupled set of linear difference equations[1]

$$x_t = A_t x_{t-1} + w_t, \qquad w_t \sim \mathcal{N}(\bar{x}_t, Q_t) \tag{1}$$

$$y_t = C_t x_t + v_t, \qquad v_t \sim \mathcal{N}(\bar{y}_t, R_t) \tag{2}$$

Here t indexes time from 1 to T, x_t is the unobserved state vector, y_t is the observation vector; w_t, v_t are Gaussian noise random variables with means $\bar{x}_t$, $\bar{y}_t$ and covariances Q_t, R_t respectively. The means may be used to model inputs, $\bar{x}_t \equiv B_t u_t$, $\bar{y}_t \equiv D_t u_t$, where u_t is a deterministic external input. A continuous time formulation has the obvious differential analog. These models have found widespread use in many areas of engineering and physics, in particular where physical systems may be well approximated by a dynamical noisy linear parameterization.

An alternative probabilistic formulation of the above equations is

$$p(x_t|x_{t-1}) = \mathcal{N}(A_t x_{t-1} + \bar{x}_t, Q_t) \tag{3}$$

$$p(y_t|x_t) = \mathcal{N}(C_t x_t + \bar{y}_t, R_t) \tag{4}$$

which define a joint Gaussian probability distribution

$$p(x_{1:T}, y_{1:T}) = \prod_{t=1}^{T} p(y_t|x_t) p(x_t|x_{t-1}) \tag{5}$$

where, by convention, $p(x_1|x_0)$ is a Gaussian distribution with mean $\bar{x}_1$ and covariance Q_1.

Given a set of observations $y_1, \ldots, y_T$, which we denote $y_{1:T}$, the two main interests are in calculating the *filtered* posterior inference $p(x_t|y_{1:t})$ and the *smoothed* posterior inference $p(x_t|y_{1:T})$ which, due to the linear Gaussian nature of the setup, are equivalent to optimal least squares estimators[1]. For smoothing we need to calculate the means $\langle x_t \rangle_{p(x_t|y_{1:T})}$ and covariances $\langle \Delta x_t \Delta x_t^T \rangle_{p(x_t|y_{1:T})}$. Throughout the paper, angled brackets $\langle f(x) \rangle$ denote averages with respect to a distribution obvious from the context. Δx_t denotes $x_t - \langle x_t \rangle$.

Existing smoothing approaches may be essentially be described as either 'Forward-Backward' methods or 'correction' methods. In the Forward-Backward methods, one may use the standard approach to inference in chain-like structures, a special case of the more general Belief Propagation algorithm which performs inference on singly-connected structures[3, 4]. This results in a set of two *independent* recursions. After the iterations are finished, the results from the forward and backward passes are combined to produce the smoothed estimate $p(x_t|y_{1:T})$. Some of the earliest work in this respect are the method of Mayne[5] and the two-filter method of Fraser and Potter[6]. Whilst very closely related, the two-filter approach does not correspond exactly to the standard Forward-Backward equations one would derive from Belief Propagation. In the two-filter method, one reverses the direction of the Filter and computes another forward pass on the reversed Filter. The method is, arguably, slightly inelegant since it requires an extra insight into how to combine these two passes into a consistent smoothed estimate. This is somewhat unnecessary, since Belief Propagation provides a natural and stable method, provided that it is implemented suitably.

In the 'correction' methods, one first runs a forward filtering method to calculate $p(x_t|y_{1:t})$, and then uses these results to work backwards in time, 'correcting' $p(x_t|y_{1:t})$ to form a smoothed estimate $p(x_t|y_{1:T})$. Classic algorithms for this are the Rauch-Tung-Striebel and fixed interval smoothers[7, 1], which may easily be applied in the case of small noise covariances.

An advantage of the Forward-Backward methods are that they correspond to well-known general methods for performing statistical inference for which deriving inference recursions for distributions

[1] In the literature sometimes the matrices A_t, C_t, Q_t are trivially temporally displaced, and the noise covariances Q_t and R_t may be structured.

which have the form of a tree (singly-connected graph) is straighforward. Furthermore, the independent form of the recursions means that they can be distributed over multiple processors, speeding up computation. In the computationally more difficult case of multiply-connected graphs, there is a growing evidence to suggest that the (strictly speaking erroneous) application of these so-called Belief Propagation methods can often form very useful approximations[8]. However, it is not obvious how either the two-filter method or the Rauch-Tung-Striebel method could be used to form approximations in this case.

The forward pass, which is common to both the Forward-Backward and correction methods, may be interpreted as updating means and covariance matrices, which avoids requiring explicit integration during the derivation. Furthermore, since one is able to work directly with moments, this will produce a form of the forward recursion which is suitable for small noise.

The standard derivation of the backpass is a little more tedious since it does not correspond to updating moments. The contribution of this paper is a technique for deriving smoothing estimates easily within the probabilistic framework and which enables one to work directly with moments during the backward pass. It is easily extend to more general singly connected structures, and we have found it useful in other smoother derivations.

Forward-Backward Equations (Belief Propagation)

For readers less familiar with the probabilistic approach, we'll briefly describe here how to perform inference on simple chain distributions[4, 8]. First, let's simplify the notation, and write the distribution as

$$p = \prod_t \phi\left(x_{t-1}, y_{t-1}, x_t, y_t\right)$$

where in the case of the Kalman-Filter, $\phi\left(x_{t-1}, y_{t-1}, x_t, y_t\right) = p(x_t|x_{t-1})p(y_t|x_t)$. Our aim is to define 'messages' ρ, λ (these correspond to the α and β messages in the Hidden Markov Model framework[9, 10]) which contain information from past observations and future observations respectively. Explicitly, we define $\rho_t(x_t)$ to represent knowledge about x_t given all information from time 1 to t, and $\lambda_t(x_t)$ to represent knowledge about state x_t given all information from the future observations from time T to time $t+1$. In the sequel, we drop the time suffix for notational clarity. The marginal inference is then given by

$$p(x_t|y_{1:T}) \propto \rho\left(x_t\right)\lambda\left(x_t\right) \tag{6}$$

Similarly, the pairwise marginal is given by

$$p(x_{t-1}, x_t|y_{1:T}) \propto \rho\left(x_{t-1}\right)\phi\left(x_{t-1}, y_{t-1}, x_t, y_t\right)\lambda\left(x_t\right) \tag{7}$$

Taking the above equation as a starting point, we can calculate the marginal from this

$$p(x_t|y_{1:T}) \propto \int_{x_{t-1}} \rho\left(x_{t-1}\right)\phi\left(x_{t-1}, y_{t-1}, x_t, y_t\right)\lambda\left(x_t\right) \tag{8}$$

Consistency with equation (6) requires (neglecting irrelevant scalings)

$$\rho\left(x_t\right)\lambda\left(x_t\right) \propto \int_{x_{t-1}} \rho\left(x_t\right)\phi\left(x_{t-1}, y_{t-1}, x_t, y_t\right)\lambda\left(x_t\right) \tag{9}$$

Similarly, we can integrate equation (7) over x_t to get the marginal at time x_{t-1}. Using then

$$p(x_{t-1}|y_{1:T}) \propto \rho\left(x_{t-1}\right)\lambda\left(x_{t-1}\right) \tag{10}$$

we arrive at the condition

$$\rho\left(x_{t-1}\right)\lambda\left(x_{t-1}\right) \propto \int_{x_t} \rho\left(x_{t-1}\right)\phi\left(x_{t-1}, y_{t-1}, x_t, y_t\right)\lambda\left(x_t\right) \tag{11}$$

Hence,

$$\text{Forward Recursion: } \rho\left(x_t\right) \propto \int_{x_{t-1}} \rho\left(x_{t-1}\right)\phi\left(x_{t-1}, y_{t-1}, x_t, y_t\right) \tag{12}$$

$$\text{Backward Recursion: } \lambda\left(x_{t-1}\right) \propto \int_{x_t} \phi\left(x_{t-1}, y_{t-1}, x_t, y_t\right)\lambda\left(x_t\right) \tag{13}$$

which are the usual definitions of the messages defined as a set of independent recursions. The extension to more general singly connected structures is straightforward and results in partially independent recursions which communicate only at branches of the tree[4].

The application of these equations in linear Gaussian state space models to perform smoothing is, mathematically speaking, trivial, since all integrals are simply Gaussians, and the messages can be represented by exponentiated quadratic forms. However, the many different, yet algebraically equivalent, implementations of the above recursions differ in their numerical stability[11].

2 Forward Pass

Whilst the derivation of the forward pass is standard, we include it here as an introduction to our approach for the backward pass. We start with the recursion

$$\rho(x_t) = \int_{x_{t-1}} p(y_t|x_t)p(x_t|x_{t-1})\rho(x_{t-1})$$

It is clear that the $\rho(x_t)$ will be Gaussian distributions, which we may write in the 'moment' form:

$$\rho(x_t) \propto \exp\left(-\frac{1}{2}\left(x_t - \tilde{f}_t\right)^T \tilde{F}_t^{-1}\left(x_t - \tilde{f}_t\right)\right) \tag{14}$$

The integrand is proportional to $p(x_t, x_{t-1}, y_t|y_{1:t-1})$, and the integral is the marginal $p(x_t, y_t|y_{1:t-1}) \propto p(x_t|y_{1:t})$. Proportionality constants are irrelevant in this context. The forward pass will then correspond to updating the moments $\tilde{f}$ and $\tilde{F}$. Hence we just need to condition the joint distribution $p(x_t, y_t|y_{1:t-1})$ on the observation y_t to find $\rho(x_t)$. The joint distribution, $p(x_t, y_t|y_{1:t-1})$ will be a Gaussian with means and covariances found easily from equations (1,2) as follows:

$$\langle x_t\rangle = A_t\left\langle x_{t-1}\right\rangle + \bar{x}_t = A_t\tilde{f}_{t-1} + \bar{x}_t, \qquad \langle y_t\rangle = \langle C_t[A_t x_{t-1} + w_{t-1}] + v_t\rangle = C_t\left(A_t\tilde{f}_{t-1} + \bar{x}_t\right) + \bar{y}_t$$

$$\Delta x_t = A_t\left[x_{t-1} - \langle x_{t-1}\rangle\right] + w_{t-1}, \qquad \Delta y_t = C_t A_t\left[x_{t-1} - \langle x_{t-1}\rangle\right] + C_t w_{t-1} + v_t$$

$$\Sigma_{xx} = \langle\Delta x_t \Delta x_t^T\rangle = A_t\tilde{F}_{t-1}A_t^T + Q_t, \qquad \Sigma_{xy} = \langle\Delta x_t\Delta y_t^T\rangle = \left(A_t\tilde{F}_{t-1}A_t^T + Q_t\right)C_t^T = \Sigma_{xx}C_t^T$$

$$\Sigma_{yy} = \langle\Delta y_t\Delta y_t^T\rangle = C\left(A_t\tilde{F}_{t-1}A_t^T + Q_t\right)C_t^T + R_t = C_t\Sigma_{xx}C_t^T + R_t$$

For a Gaussian $p(x, y)$ with covariances $\Sigma_{xx}, \Sigma_{xy}, \Sigma_{yy}$ and means μ_x, μ_y, the distribution $p(x|y)$ has mean $\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}\left(y - \mu_y\right)$ and covariance $\Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$. Hence $p(x_t|y_{1:t})$ has covariance and mean

$$\tilde{F}_t = P_t - P_t C_t^T(C_t P_t C_t^T + R)^{-1}C_t P_t^T = (I - H_t C_t)P_t \tag{15}$$

$$\tilde{f}_t = (I - H_t C_t)A_t\tilde{f}_{t-1} + H_t\left(y_t - C_t\bar{x}_t - \bar{y}_t\right) \tag{16}$$

where $H_t = P_t C_t^T(C_t P_t C_t^T + R_t)^{-1}$, $P_t \equiv \Sigma_{xx} = A_t\tilde{F}_{t-1}A_t^T + Q_t$. Equations (15,16) form the forward pass of the Kalman filter. The equations may conversely be written in the canonical form under reparameterization $F = \tilde{F}^{-1}$,$f = \tilde{F}^{-1}\tilde{f}$, for which the recursions are expected to be more stable in the regime of large noise.

Algorithm 1 The Kalman Filter : Forward Pass in Moment Form

procedure KalmanForwardMoment($y_{1:T}$,$A_{2:T}$,$C_{1:T}$,$Q_{1:T}$,$R_{1:T}$,$\bar{x}_{1:T}$,$\bar{y}_{1:T}$)
  $\tilde{F}_0 \leftarrow 0,\ \tilde{f}_0 \leftarrow 0$
  for $t \leftarrow 1, T$ **do**
    $P_t \leftarrow A_t \tilde{F}_{t-1} A_t^T + Q_t$
    $H_t \leftarrow P_t C_t^T \left(C_t P_t C_t^T + R_t\right)^{-1}$
    $\tilde{F}_t \leftarrow (I - H_t C_t) P_t$
    $\tilde{f}_t \leftarrow (I - H_t C_t) A_t \tilde{f}_{t-1} + H_t (y_t - C_t \bar{x}_t - \bar{y}_t)$
  end for
end procedure

3 Backward Pass

The Belief Propagation equations are

$$\lambda(x_{t-1}) = \int_{x_t} p(y_t|x_t) p(x_t|x_{t-1}) \lambda(x_t)$$

What is disadvantageous about this equation is that it does not represent a probability distribution over x_{t-1}, but rather a *conditional* distribution $\lambda(x_t) \propto p(y_{t+1:T} \,|\, x_t)$. For this reason, we are forced to use the canonical representation for the λ messages:

$$\lambda(x_t) \propto \exp\left(-\frac{1}{2}\left(x_t^T G_t x_t - 2x_t^T g_t\right)\right)$$

where G is non positive-definite symmetric matrix. A straightforward evaluation of the integral results in the λ recursion (for the zero mean case $\bar{x} \equiv 0$,$\bar{y} \equiv 0$)

$$G_{t-1} = A_t^T \left(Q_t^{-1} - Q_t^{-1} L_t^{-1} Q_t^{-1}\right) A_t \tag{17}$$

$$g_{t-1} = A_t^T Q_t^{-1} L_t^{-1} \left(C_t^T R_t^{-1} + g_t\right) \tag{18}$$

$$L_t = Q_t^{-1} + C_t^T R_t^{-1} C_t + G_t \tag{19}$$

However, this form is inappropriate for small noise covariances. Whilst it is indeed possible to find a rearrangement by algebraic manipulation which avoids explicit use of Q^{-1} and R^{-1} (see appendix), this lacks insight and generality. As an alternative, we present below a method which is somewhat general, and will always result in the form of the backpass appropriate for the small noise case in related models. Rather than the creative use of Woodbury algebraic manipulations for each new model, which may resulting in some head-scratching, the following is a simple handle-turning method which will lead to the appropriate small-noise form of the recursion since it works directly with moments.

The Auxiliary Variable Trick

The trick is to introduce an 'auxiliary' variable a_t that, when set to a certain state, encodes the same information as the $\lambda(x_t)$ message. There are different ways to do this, but the approach taken here is perhaps the most obvious. We define a new variable a_t by :

$$a_t = G_t x_t + z_t, \qquad z_t \sim \mathcal{N}(0, G_t) \tag{20}$$

where the noise z_t is Gaussian, with zero mean and covariance G_t, thus[2]

$$p(a_t \,|\, x_t) \propto \exp\left(-\frac{1}{2}(a_t - G_t x_t)^T G_t^{-1} (a_t - G_t x_t)\right)$$

[2]Whilst the expression contains the expression G^{-1} which is formally not invertible, if we expand the quadratic form, only irrelevant proportionality constants contain G^{-1}.

Then:

$$\lambda(x_{t-1}) \propto \int_{x_t} p(y_t|x_t)p(x_t|x_{t-1})p(a_t|x_t)|_{a_t=g_t}$$

The advantage of this expression is that this can be regarded as the marginal of a joint probability,

$$\lambda(x_{t-1}) \propto \int_{x_t} p(y_t, x_t, a_t|x_{t-1})|_{a_t=g_t} = p(y_t, a_t|x_{t-1})|_{a_t=g_t}$$

Since all the noise distributions are Gaussians and transitions are linear, the joint distribution $p(y_t, a_t|x_{t-1})$ will also be Gaussian, with means and covariances

$$p(y_t, a_t \mid x_{t-1}) \propto \exp\left\{ -\frac{1}{2} \begin{pmatrix} a_t - \langle a_t\rangle \\ y_t - \langle y_t\rangle \end{pmatrix}^T \begin{pmatrix} \Sigma_{aa} & \Sigma_{ay} \\ \Sigma_{ay}^T & \Sigma_{yy} \end{pmatrix}^{-1} \begin{pmatrix} a_t - \langle a_t\rangle \\ y_t - \langle y_t\rangle \end{pmatrix} \right\} \tag{21}$$

$$\begin{aligned}
&\langle a_t\rangle = G_t\langle x_t\rangle = G_t\left(A_t x_{t-1} + \bar{x}_t\right), \qquad \langle y_t\rangle = C_t\langle x_t\rangle = C_t\left(A_t x_{t-1} + \bar{x}_t\right) + \bar{y}_t \\
&\Delta a_t = a_t - \langle a_t\rangle = G_t w_t + z_t, \qquad \Delta y_t = y_t - \langle y_t\rangle = C_t w_t + v_t \\
&\Sigma_{aa} = \langle \Delta a_t \Delta a_t^T\rangle = G_t Q_t G_t^T + G_t, \qquad \Sigma_{vv} = \langle \Delta y_t \Delta y_t^T\rangle = C_t Q_t C_t^T + R_t \\
&\Sigma_{av} = \langle \Delta a_t \Delta y_t^T\rangle = G_t Q_t C_t^T
\end{aligned}$$

In these expressions, the noise covariances appear, but not their inverses. After setting $a_t = g_t$, our interest is in finding the functional dependence of equation (21) on x_{t-1}, since this is what defines the backward $\lambda(x_{t-1})$ message. Using $\begin{pmatrix} D_{aa} & D_{ay} \\ D_{ya} & D_{yy} \end{pmatrix} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ay} \\ \Sigma_{ya} & \Sigma_{yy} \end{pmatrix}^{-1}$, this dependence is given by

$$\exp\left(-\frac{1}{2}\left(x_{t-1}^T G_{t-1} x_{t-1} - 2x_{t-1}^T g_{t-1}\right)\right)$$

where we ignored unnecessary constants, with

$$\begin{aligned}
G_{t-1} &= A_t^T\left(G_t D_{aa} G_t + G_t^T D_{ay} C_t + C_t^T D_{ay}^T G_t + C_t^T D_{yy} C_t\right) A_t \\
g_{t-1} &= A_t^T\left(G_t D_{aa} \hat{g}_t + C_t^T D_{ya} \hat{g}_t + G_t D_{ay} \hat{y}_t + C_t^T D_{yy} \hat{y}_t\right) \\
\hat{g}_t &\equiv g_t - G_t \bar{x}_t, \qquad \hat{y}_t \equiv y_t - C_t \bar{x}_t - \bar{y}_t
\end{aligned}$$

Using the inverse of a partitioned matrix:

$$\begin{aligned}
&D_{ay} = -D_{aa}(\Sigma_{ay}\Sigma_{yy}^{-1}), \qquad D_{yy} = \Sigma_{yy}^{-1} + (\Sigma_{yy}^{-1}\Sigma_{ay}^T) D_{aa} (\Sigma_{ay}\Sigma_{yy}^{-1}) \\
&\Sigma_{ay}\Sigma_{yy}^{-1} = G_t \underbrace{Q_t C_t^T (C Q_t C_t^T + R_t)^{-1}}_{\equiv K_t}
\end{aligned}$$

$$\begin{aligned}
D_{aa} &= (\Sigma_{aa} - \Sigma_{ay}\Sigma_{yy}^{-1}\Sigma_{ay}^T)^{-1} = \left(G_t Q_t G_t + G_t - G_t K C_t \Sigma_h G_t\right)^{-1} \\
&= G_t^{-1}\Big[G_t \underbrace{(I - K_t C_t)}_{\equiv E_t} Q_t + I\Big]^{-1} = G_t^{-1}\left(G_t E_t Q_t + I\right)^{-1}
\end{aligned} \tag{22}$$

we get:

$$G_{t-1} = A_t^T E_t^T \left(G_t E_t Q_t + I\right)^{-1} G_t E_t A_t + A_t^T C_t^T \left(C_t Q_t C_t^T + R_t\right)^{-1} C_t A_t \tag{23}$$

$$g_{t-1} = A_t^T\Big[E_t^T\left(G_t E_t Q_t + I\right)^{-1}\left(\hat{g}_t - G_t K_t \hat{y}_t\right) + C_t^T\left(C_t Q_t C_t^T + R_t\right)^{-1}\hat{y}_t\Big] \tag{24}$$

Equations (23,24) form the backward pass of the Kalman filter, without explicit reference to inverse noise covariances. That these equations are algebraically equivalent to equations (17, 18) is shown in the appendix. In the algorithm presented below, we eliminated the variable K_t to make the algorithm look a little cleaner.

Algorithm 2 The Kalman Filter : Backward Pass in Stable Canonical Form

procedure KALMANBACKWARDSMALLNOISE($y_{1:T}$,$A_{2:T}$,$B_{1:T}$,$Q_{1:T}$,$R_{1:T}$,$\bar{x}_{1:T}$,$\bar{y}_{1:T}$)
  $G_T \leftarrow 0$, $g_T \leftarrow 0$
  for $t \leftarrow T, 2$ **do**
    $N_t \leftarrow \left(C_t Q_t C_t^T + R_t\right)^{-1}$
    $E_t \leftarrow I - Q_t C_t^T N_t C_t$
    $\hat{L}_t \leftarrow (G_t E_t Q_t + I)^{-1}$
    $\hat{g}_t \leftarrow g_t - G_t \bar{x}_t$
    $\hat{y}_t \leftarrow y_t - C_t \bar{x}_t - \bar{y}_t$
    $G_{t-1} \leftarrow A_t^T \left(E_t^T \hat{L}_t G_t E_t + C_t^T N_t C_t\right) A_t$
    $g_{t-1} \leftarrow A_t^T \left(E_t^T \hat{L}_t \left(\hat{g}_t - G_t Q_t C_t^T N_t \hat{y}_t\right) + C_t^T N_t \hat{y}_t\right)$
  end for
end procedure

The posterior inference $p(x_t|y_{1:T})$ has mean $(F_t + G_t)^{-1}(f_t + g_t)$, and covariance $(F_t + G_t)^{-1}$, where $F_t \equiv \tilde{F}_t^{-1}$, $f_t \equiv \tilde{F}_t^{-1}\tilde{f}_t$.

4 Discussion

Our interest here was in the derivation of Kalman smoothers using the Forward-Backward approach, a special case of Belief Propagation. Deriving the Forward Pass is easy since it corresponds to updating moments of Gaussians. In the Backpass, however, this is not the case, and results in a version inappropriate for small noise covariances. We introduced a technique that enables one to directly work with moments during the Backpass and automatically results in a recursion appropriate for small noise, obviating the need for creative algebraic manipulations. It is not clear how other approaches such as the two-filter, Rung-Tauch-Striebel and fixed interval smoother methods could be used to form useful approximations in more complex multiply-connected structures, which motivates interest in the Forward-Backward approach, in addition to its ease of parallelization.

In summary, the auxiliary variable trick is a simple technique for deriving smoothing recursions, and we have found it useful in other models. Didactically, if nothing else, it somewhat simplifies and makes more general the analysis of linear Gaussian state space models.

Appendix

Here we show (for the zero-mean noise case) how to construct the new recursions starting from those derived by the standard integration method, equations (17,18,19).

Construction for G

We begin with the original equations (17,19),

$$L - G = Q^{-1} + C^T R^{-1} C \equiv S^{-1}$$

where using the Woodbury formula,

$$\left(A + L^T C R\right)^{-1} = A^{-1} - A^{-1} L^T \left(C^{-1} + R A^{-1} L^T\right)^{-1} R A^{-1}$$

we can write $S = Q - QC^T NCQ \equiv EQ$. Then

$$\begin{aligned} Q - L^{-1} &= Q - \left(S^{-1} + G\right)^{-1} = Q - S\left(I + GS\right)^{-1}\left(\left(I + GS\right) - GS\right) \\ &= \underbrace{Q - S}_{QC^T NCQ} + S\left(I + GS\right)^{-1} GS \end{aligned}$$

Hence

$$Q^{-1} - Q^{-1} L^{-1} Q^{-1} = Q^{-1}\left(QC^T NCQ\right) Q^{-1} + \underbrace{Q^{-1} S}_{E^T} \hat{L} G \underbrace{SQ^{-1}}_{E}$$

or

$$Q^{-1} - Q^{-1} L^{-1} Q^{-1} = C^T NC + E^T \hat{L} G E$$

which is the form of the recursion for G in equation (23) when pre-multiplied by A and postmultiplied by A^T.

Construction for g

The update equations (18,19) for g contains two contributions : one from g, and one from y.

$$Q^{-1} L^{-1} = Q^{-1}\left(S^{-1} + G\right)^{-1} = Q^{-1} S\left(I + GS\right)^{-1} = E^T \hat{L} \tag{25}$$

which shows that the g coefficients of equation (18) and equation (24) match. We'll now look at the coefficient for y :

$$Q^{-1} L^{-1} C^T R^{-1} = Q^{-1}\left(I - L^{-1} G\right)\left(L - G\right)^{-1} C^T R^{-1}$$

Now, consider

$$\begin{aligned} \left(L - G\right)^{-1} C^T R^{-1} &= SC^T R^{-1} = \left(Q - QC^T NCQ\right) C^T R^{-1} \\ &= QC^T N\left(N^{-1} - CQC^T\right) R^{-1} = QC^T N \end{aligned}$$

Hence

$$\begin{aligned} &Q^{-1} L^{-1} C^T R^{-1} = Q^{-1}\left(I - L^{-1} G\right) QC^T N \\ &= C^T N - Q^{-1} L^{-1} GQC^T N = C^T N - E^T \hat{L} GQC^T N \end{aligned}$$

which shows that the coefficients of y in (18,24) match.

Acknowledgment

I would especially like to thank Peter Sollich for deriving the algebraic equivalences of the two forms of the backpass. Many thanks also to Ali Taylan Cemgil, Bertrand Mesot and Silvia Chiappa for useful discussions.

References

[1] J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.

[2] Y. Bar-Shalom and Xiao-Rong Li. *Estimation and Tracking : Principles, Techniques and Software*. Artech House, Norwood, MA, 1998.

[3] S. L. Lauritzen. *Graphical Models.* Oxford University Press, 1996.

[4] F. V. Jensen. *Bayesian Networks and Decision Graphs.* Springer Verlag, 2001.

[5] D. Q. Mayne. A Solution of the Smoothing Problem for Linear Dynamical Systems. *Automatica*, 4:73–92, 1966.

[6] D. C. Fraser and J. E. Potter. The Optimum Linear Smoother as a Combination of Two optimal Linear Filters. *IEEE Transactions of Automatic Control*, 14(4):387–390, 1969.

[7] H. E. Rauch, G. Tung, and C. T. Striebel. Maximum Likelihood estimates of linear dynamic systems. *American Institute of Aeronautics and Astronautics Journal (AIAAJ)*, 3(8):1445–1450, 1965.

[8] M. A. Srinvas and R. J. McEliece. The Generalised Distributive Law. *IEEE Transactions of Information Theory*, 46(2):325–343, 2000.

[9] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2), 1989.

[10] B. Deylon. Remarks on Linear and Nonlinear Filtering. *IEEE Transactions on Information Theory*, 41(1):317–322, 1995.

[11] Verhaegen. M. and P. Van Dooren. Numerical Aspects of Different Kalman Filter Implementations. *IEEE Transactions of Automatic Control*, 31(10):907–917, 1986.