Kalman Filtering: A PGM derivation

SCJ Robertson

21st May, 2016

Introduction

The purpose of this document is to derive the Kalman Filter algorithm from a PGM perspective. It is shown that when the Kalman Filter is represented as a Junction Tree the standard algorithm is hiding in plain sight and just needs to be extracted from the results of the Integral-Sum message passing algorithm.

1. Linear Gaussian Systems

The following three conditions are sufficient, in addition to the Markov assumption, to ensure that the belief, $p(\boldsymbol{x}_t)$, of the system is always Gaussian [8, 9]:

1. The state transition probability $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{u}_t)$ is a linear function in its arguments, with added Gaussian noise.

$$\boldsymbol{x}_t = A_t\boldsymbol{x}_{t-1} + B_t\boldsymbol{u}_t + \boldsymbol{\epsilon}_t \tag{1.1}$$

   A_t and B_t are $n \times n$ and $m \times n$ matrices respectively, with n being the dimension of the sate vector. $\boldsymbol{\epsilon}_t$ is a Gaussian random vector which accounts for the uncertainty introduced during state transition, it is zero mean with a covariance matrix R_t. The state transition probability is now given as:

$$\begin{aligned}
&p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{u}_t) = \mathcal{N}\left(\boldsymbol{x}_t|A\boldsymbol{x}_{t-1} + B\boldsymbol{u}_t, R_t\right) \\
&= \frac{1}{(2\pi)^{n/2}|R|^{1/2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}_t - A_t\boldsymbol{x}_{t-1} + B\boldsymbol{u}_t\right)^T R_t^{-1}\left(\boldsymbol{x}_t - A_t\boldsymbol{x}_{t-1} + B\boldsymbol{u}_t\right)\right\}
\end{aligned} \tag{1.2}$$

2. The measurement probability, $p(\boldsymbol{z}_t|\boldsymbol{x}_t)$, must also be linear in its arguments, with added Gaussian Noise:

$$\boldsymbol{z}_t = C_t\boldsymbol{x}_t + \boldsymbol{\delta}_t \tag{1.3}$$

   C_t is $k \times n$ matrix, $\boldsymbol{z}_t$ being k dimensional. $\boldsymbol{\delta}_t$ is zero mean Gaussian noise with a covariance of Q_t. $p(\boldsymbol{z}_t|\boldsymbol{x}_t)$ is now defined as:

$$\begin{aligned} p(\boldsymbol{z}_t|\boldsymbol{x}_t) &= \mathcal{N}(\boldsymbol{z}_t|C_t\boldsymbol{x}_t, Q_t) \\ &= \frac{1}{(2\pi)^{k/2}|Q|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_t - C_t\boldsymbol{x}_t)^T Q_t^{-1}(\boldsymbol{z}_t - C_t\boldsymbol{x}_t)\right\} \end{aligned} \tag{1.4}$$

3. The belief about the initial state, $p(\boldsymbol{x}_0)$, must be normally distributed:

$$p(\boldsymbol{z}_t|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_0|\boldsymbol{\mu}_t, \Sigma_t)$$

Remark 1 *This entire section was shamelessly stolen from [9], but it's necessary to restate all of this to establish the ground rules and introduce some notation. The notation quickly becomes an unbearable mess from this point onwards.*

1.1 Bayes Net Representation

Definition 1 (Bayesian Network) *A Bayesian Network is a distribution of the form:*

$$p(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) = p(\boldsymbol{x}_0^N) = \prod_{i=1}^{N} p(x_i|pa(\boldsymbol{x}_i)) \tag{1.5}$$

where $pa(x_i)$ are the ***parental*** *variables of x_i. It is represented naturally as a directed graph, with a directed edge from the parent to its child. The j^{th} node corresponds to the factor $p(\boldsymbol{x}_j|pa(\boldsymbol{x}_j))$ [1, 5].*

The Kalman Filter can be represented as a joint density function of all random variables $\boldsymbol{x}_0^t$ and $\boldsymbol{z}_1^t$, without any loss of generality[1]. This allows the following factorisation:

$$\begin{aligned} p(\boldsymbol{x}_0^t, \boldsymbol{z}_1^t, \boldsymbol{u}_1^t) &= p(\boldsymbol{z}_t|\boldsymbol{x}_t, \boldsymbol{z}_1^{t-1}, \boldsymbol{u}_1^t)\, p(\boldsymbol{x}_0^t, \boldsymbol{z}_1^{t-1}, \boldsymbol{u}_1^t) \\ &= p(\boldsymbol{z}_t|\boldsymbol{x}_t, \boldsymbol{z}_1^{t-1}, \boldsymbol{u}_1^t)\, p(\boldsymbol{x}_0^t|\boldsymbol{x}_0^{t-1}, \boldsymbol{z}_1^{t-1}, \boldsymbol{u}_1^t)\, p(\boldsymbol{x}_0^{t-1}, \boldsymbol{z}_1^{t-1}, \boldsymbol{u}_1^{t-1}) \end{aligned} \tag{1.6}$$

Applying the Markov assumption to Equation 1.6 and then following the recursion to its conclusion, results in:

$$p(\boldsymbol{x}_0^t, \boldsymbol{z}_1^t, \boldsymbol{u}_1^t) = p(\boldsymbol{x}_0) \prod_{i=1}^{t} p(\boldsymbol{z}_i|\boldsymbol{x}_i)\, p(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}, \boldsymbol{u}_t) \tag{1.7}$$

The form of Equation 1.7 is exactly that of Equation 1.5 allowing the Kalman Filter to be represented naturally as a Bayes Net, as seen in Figure 1.1.

[1] Okay, I'm not sure if this is true, but I've always wanted to say it.

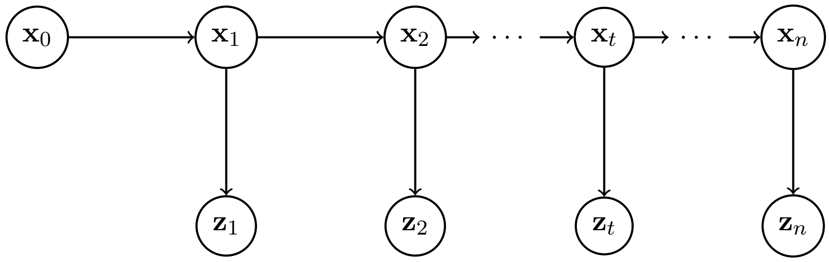

Figure 1.1: A Bayes Net for a Kalman filter. For brevity's sake, the control vector, $\boldsymbol{u}_t$ has been omitted as it is deterministic and known in each state.

Remark 2 *The representation of Equation 1.7 is misleading. The control vector,* $\boldsymbol{u}_t$*, is not a random variable, it is deterministic and known in each state.*

1.2 Junction Tree Representation

The Junction Tree Algorithm is a form of tree decomposition which seeks to eliminate cycles in a graph by clustering them into a single node. With the cycles removed it is possible to perform exact inference, avoiding any iterative message passing schemes.

Remark 3 *I was going to define clusters graphs, the generalised running intersection property and then arrive at clique trees, but it seemed daunting. Also pointless, considering I'm the only person who will ever read this and any definitions I give will just be rehashed versions of those given by Koller and Barber.*

1.2.1 Junction Tree Construction: The HUGIN Algorithm

The HUGIN algorithm provides a procedure for constructing a Junction tree from some general graph [4]:

1. If the graph if directed, moralize it. Moralization will produce the equivalent undirected form of the graph, preserving the underlying probabilistic dependencies.
2. Triangulate the graph using variable elimination, this will produce the induced graph. Elimination ordering determines the density of the induced graph, an ordering is sought which will introduce the least amount of fill edges.
3. Identify all maximal cliques in the induced graph, the variables in cliques are then assigned to this *supernode.*
4. Allocate each potential from the original graph to exactly one of these *super* nodes. The *supernode* is simply a product of all of its associated potentials.
5. Find a maximal spanning tree over all the *supernodes.*

Application to the Kalman Filter

The Kalman Filter's Bayes Net representation is already a tree, therefore its decomposition is trivial. Moralization introduces no new edges as all child nodes only have a single parent. The variable elimination orderings are also trivial, the leaf nodes can eliminated in any order and will also introduce no fill edges into the induced graph. The remaining chain structure only has two possible elimination orderings, forwards or backwards, both are perfect and introduce

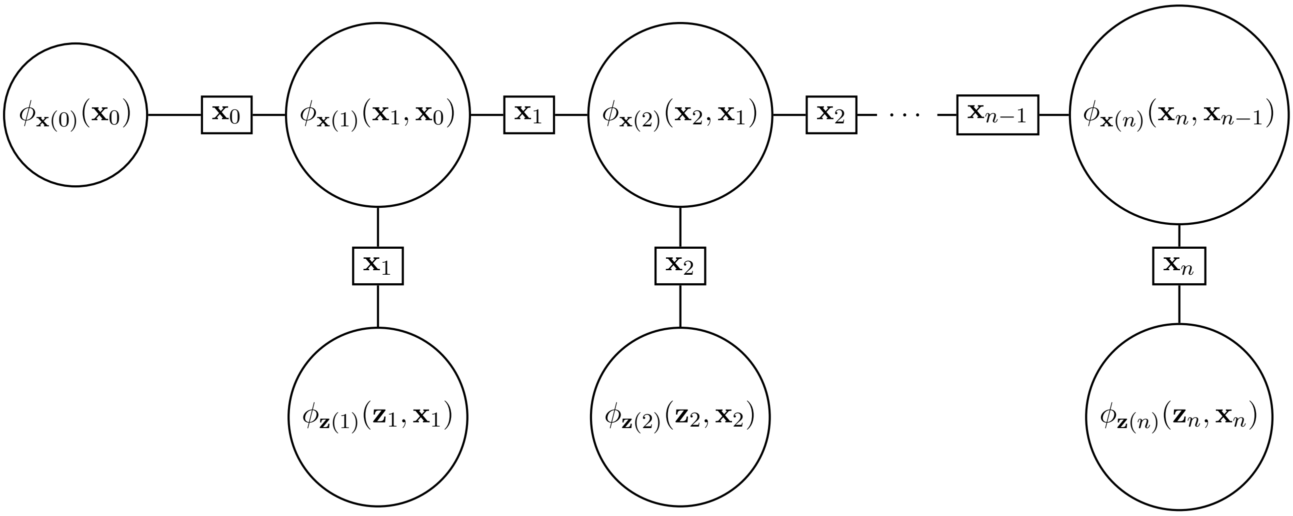

Figure 1.2: The Junction Tree resulting from the Bayes Net in Figure 1.1.

no fill edges. All maximal cliques comprise of a adjacent parings of nodes, which only need to assigned a single potential. The maximal spanning tree is the existing tree as sepsets are single variable sets, $\boldsymbol{S}_{i,j} = \boldsymbol{C}_i \cap \boldsymbol{C}_j = \{\boldsymbol{x}_i\}$.

The Junction Tree equivalent of Figure 1.1 can be seen in Figure 1.2.

1.2.2 Message Passing: Integral-Product Algorithm

Definition 2 (Integral-Product Message Passing) *In a cluster graph, $\mathcal{T}$, defined for a set of continuous factors Φ over $\mathcal{X}$, the message from a cluster $\boldsymbol{C}_i$ to an adjacent cluster $\boldsymbol{C}_j$ is computed using the* ***integral-product message passing*** *computation:*

$$\delta_{i\to j} = \int \psi_i \left(\prod_{k \in (Nb_i - \{j\})} \delta_{k\to i} \right) d\{\boldsymbol{C}_i - \boldsymbol{S}_{i,j}\} \tag{1.8}$$

Here ψ_i is the potential assigned to $\boldsymbol{C}_i$ and Nb_i are all clusters adjacent to it [3, 6, 7].

From Definition 2 it can be seen that the Junction Tree algorithm does not necessarily simplify inference, with its larger nodes the work of iterative message passing is simply shifted to this integral.

2. Kalman Filtering

The current state estimate, $p(\boldsymbol{x}_t)$, is equivalent to the message $\delta_{\phi_{\boldsymbol{x}_t}\to\phi_{\boldsymbol{x}_{t+1}}}(\boldsymbol{x}_t)$ from $\phi_{\boldsymbol{x}_t}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$ to $\phi_{\boldsymbol{x}_{t+1}}(\boldsymbol{x}_t, \boldsymbol{x}_t)$. Figure 2.1 shows the message flow which results in:

$$\begin{aligned} \delta_{\boldsymbol{x}(t)\to\boldsymbol{x}(t+1)}(\boldsymbol{x}_t) &= \int \phi_{\boldsymbol{x}(t)}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) \delta_{\boldsymbol{z}(t)\to\boldsymbol{x}(t)}(\boldsymbol{x}_t) \delta_{\boldsymbol{x}(t-1)\to\boldsymbol{x}(t)}(\boldsymbol{x}_{t-1}) d\boldsymbol{x}_{t-1} \\ &= \underbrace{\delta_{\boldsymbol{z}(t)\to\boldsymbol{x}(t)}(\boldsymbol{x}_t)}_{\textit{Measurement update}} \underbrace{\int \phi_{\boldsymbol{x}(t)}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) \delta_{\boldsymbol{x}(t-1)\to\boldsymbol{x}(t)}(\boldsymbol{x}_{t-1}) d\boldsymbol{x}_{t-1}}_{\textit{Prediction}} \end{aligned} \tag{2.1}$$

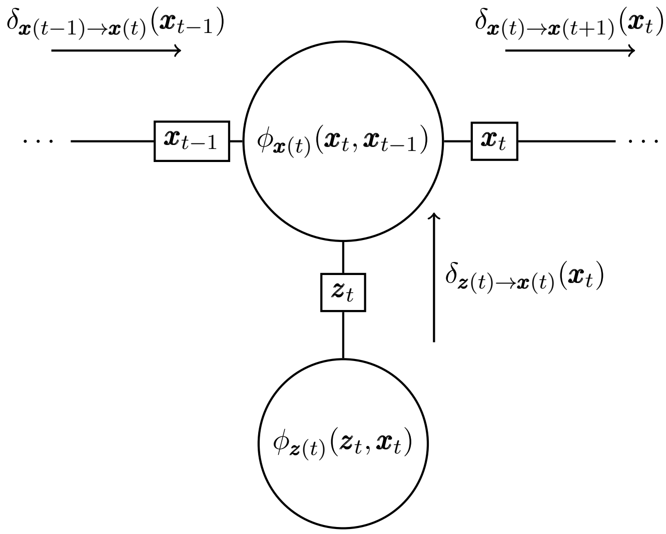

Figure 2.1: Message passing in a Kalman Filter.

Remark 4 *The notation in Equation 2.1 is unwieldy, but I couldn't think of anything else that wasn't ambiguous.*

2.1 Part 1: Prediction

Basically, evaluating this equation:

$$\Psi(\boldsymbol{x}_t) = \int \phi_{\boldsymbol{x}(t)}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})\delta_{\boldsymbol{x}(t-1)\to\boldsymbol{x}(t)}(\boldsymbol{x}_{t-1})d\boldsymbol{x}_{t-1} \tag{2.2}$$

2.1.1 Canonical Form Representation

Canonical form representation is consistent with [7].

The Initial Belief, $\phi_{\boldsymbol{x}(t)}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$

The potential, $\phi_{\boldsymbol{x}(t)}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$, is the CPD:

$$\begin{aligned}
&\phi_{\boldsymbol{x}(t)}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t | A_t\boldsymbol{x}_{t-1} + B_t\boldsymbol{x}_{t-1}, R_t\right) \\
&= \frac{1}{(2\pi)^{n/2}|R_t|^{1/2}} \exp\{-\frac{1}{2}\left(\boldsymbol{x}_t - A_t\boldsymbol{x}_{t-1} - B_t\boldsymbol{x}_{t-1}\right)^T R_t^{-1}\left(\boldsymbol{x}_t - A_t\boldsymbol{x}_{t-1} - B_t\boldsymbol{x}_{t-1}\right)\}
\end{aligned} \tag{2.3}$$

It can be rearranged as a joint density function as follows:

$$\begin{aligned}
&-\frac{1}{2}\left(\boldsymbol{x}_t - A_t\boldsymbol{x}_{t-1} - B_t\boldsymbol{x}_{t-1}\right)^T R_t^{-1}\left(\boldsymbol{x}_t - A_t\boldsymbol{x}_{t-1} - B_t\boldsymbol{x}_{t-1}\right) \\
&= -\frac{1}{2}\begin{bmatrix}(\boldsymbol{x}_t - B_t\boldsymbol{u}_t)^T & \boldsymbol{x}_{t-1}^T\end{bmatrix}\begin{bmatrix} R_t^{-1} & -R_t^{-1}A_t \\ -A_t^T R_t^{-1} & A_t^T R_t^{-1} A_t\end{bmatrix}\begin{bmatrix}(\boldsymbol{x}_t - B_t\boldsymbol{u}_t) \\ \boldsymbol{x}_{t-1}\end{bmatrix} \\
&= -\frac{1}{2}\left(\boldsymbol{X}_t - \boldsymbol{M}_t\right)^T P_t\left(\boldsymbol{X}_t - \boldsymbol{M}_t\right)
\end{aligned} \tag{2.4}$$

Where,

$$\boldsymbol{X}_t = \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{x}_{t-1} \end{bmatrix} \tag{2.5}$$

$$\boldsymbol{M}_t = \begin{bmatrix} B_t \boldsymbol{u}_t \\ \mathbf{0} \end{bmatrix} \tag{2.6}$$

$$P_t = \begin{bmatrix} R_t^{-1} & -R_t^{-1} A_t \\ -A_t^T R_t^{-1} & A_t^T R_t^{-1} A_t \end{bmatrix} \tag{2.7}$$

The joint density can now be easily represented in canonical form:

$$\begin{aligned} \phi_{\boldsymbol{x}(t)}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) &= \mathcal{N}\left(\boldsymbol{X} | \boldsymbol{M}, P\right) \\ &= \mathcal{C}\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}; P_t, \boldsymbol{h}_t, g_t\right) \end{aligned} \tag{2.8}$$

Where,

$$\boldsymbol{h}_t = P_t \boldsymbol{M}_t \tag{2.9}$$

$$g_t = -\frac{1}{2} \boldsymbol{M}_t^T P_t \boldsymbol{M}_t - \ln\left\{(2\pi)^{n/2} |R_t|^{1/2}\right\} \tag{2.10}$$

Remark 5 *I introduced $\boldsymbol{X}_t$ just to show the new arrangement was in quadratic form, I find it easier to show full scope in canonical form so I don't lose track of anything.*

The Recursive Belief, $\delta_{\boldsymbol{x}(t-1)\to\boldsymbol{x}(t)}(\boldsymbol{x}_{t-1})$

$\delta_{\boldsymbol{x}(t-1)\to\boldsymbol{x}(t)}(\boldsymbol{x}_{t-1})$ is some unknown distribution which can be represented generally in canonical form:

$$\delta_{\boldsymbol{x}(t-1)\to\boldsymbol{x}(t)}(\boldsymbol{x}_{t-1}) = \mathcal{C}\left(\boldsymbol{x}_{t-1}; P_{t-1}, \boldsymbol{h}_t, g_{t-1}\right) \tag{2.11}$$

Where,

$$P_{t-1} = \Sigma_{t-1}^{-1} \tag{2.12}$$

$$\mathbf{h}_{t-1} = \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1} \tag{2.13}$$

$$g_{t-1} = -\frac{1}{2} \boldsymbol{\mu}^T \Sigma_{t-1}^{-1} \boldsymbol{\mu} - \ln\left\{(2\pi)^{n/2} |\Sigma_{t-1}|^{1/2}\right\} \tag{2.14}$$

Since $\delta_{\boldsymbol{x}(t-1)\to\boldsymbol{x}(t)}(\boldsymbol{x}_{t-1})$'s scope is a subset of $\phi_{\boldsymbol{x}(t)}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$'s, its variables must be augmented with zeros:

$$P'_{t-1} = \begin{bmatrix} 0 & -0 \\ 0 & \Sigma_{t-1}^{-1} \end{bmatrix} \tag{2.15}$$

$$\mathbf{h}'_{t-1} = \begin{bmatrix} \mathbf{0} \\ \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1} \end{bmatrix} \tag{2.16}$$

2.1.2 Belief Update

$$\begin{aligned} \phi_{\boldsymbol{x}(t)}(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}) \delta_{\boldsymbol{x}(t-1)\to\boldsymbol{x}(t)}(\boldsymbol{x}_{t-1}) &= \mathcal{C}\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}; P_t, \boldsymbol{h}_t, g_t\right) \cdot \mathcal{C}\left(\boldsymbol{x}_{t-1}; P'_{t-1}, \boldsymbol{h}'_t, g_{t-1}\right) \\ &= \mathcal{C}\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}; P_t + P'_{t-1}, \boldsymbol{h}_t + \boldsymbol{h}'_{t-1}, g_t + g_{t-1}\right) \\ &= \mathcal{C}\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}; \hat{P}_t, \hat{\boldsymbol{h}}_t, \hat{g}_t\right) \end{aligned} \tag{2.17}$$

$$
\begin{aligned}
\hat{P}_t &= P_t + P'_{t-1} \\
&= \begin{bmatrix} R_t^{-1} & -R_t^{-1}A_t \\ -A_t^T R_t^{-1} & A_t^T R_t^{-1} A_t \end{bmatrix} + \begin{bmatrix} 0 & -0 \\ 0 & \Sigma_{t-1}^{-1} \end{bmatrix} \\
&= \begin{bmatrix} R_t^{-1} & -R_t^{-1}A_t \\ -A_t^T R_t^{-1} & A_t^T R_t^{-1} A_t + \Sigma_{t-1}^{-1} \end{bmatrix} && (2.18) \\
\hat{\boldsymbol{h}}_t &= \boldsymbol{h}_t + \boldsymbol{h}'_{t-1} \\
&= \begin{bmatrix} R_t^{-1} B_t \boldsymbol{u}_t \\ -A_t^T R_t^{-1} B_t \boldsymbol{u}_t \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1} \end{bmatrix} \\
&= \begin{bmatrix} R_t^{-1} B_t \boldsymbol{u}_t \\ -A_t^T R_t^{-1} B_t \boldsymbol{u}_t + \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1} \end{bmatrix} && (2.19) \\
\hat{g}_t &= g_t + g_{t-1} \\
&= -\frac{1}{2} \boldsymbol{M}_t^T P \boldsymbol{M}_t - \ln\left\{ (2\pi)^{n/2} |R_t|^{1/2} \right\} + \boldsymbol{\mu}^T \Sigma_{t-1}^{-1} \boldsymbol{\mu} - \ln\left\{ (2\pi)^{n/2} |\Sigma_{t-1}|^{1/2} \right\} && (2.20)
\end{aligned}
$$

Marginalisation

Evaluating Equation 2.1 according to [7]:

$$
\begin{aligned}
\widetilde{P}_t &= R_t^{-1} - \left(A_t^T R_t^{-1} \right)^T \left(A_t^T R_t^{-1} A_t + \Sigma_{t-1}^{-1} \right)^{-1} \left(A_t^T R_t^{-1} \right) && (2.21) \\
\widetilde{\boldsymbol{h}}_t &= R_t^{-1} B_t \boldsymbol{u}_t + R_t^{-1} A_t \left(A_t^T R_t^{-1} A_t + \Sigma_{t-1}^{-1} \right)^{-1} \left(-A_t^T R_t^{-1} B_t \boldsymbol{u}_t + \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1} \right) && (2.22) \\
\widetilde{g}_t &= \hat{g}_t - \frac{1}{2} \left(-A_t^T R_t^{-1} B_t \boldsymbol{u}_t + \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1} \right)^T \left(A_t^T R_t^{-1} A_t + \Sigma_{t-1}^{-1} \right)^{-1} \left(-A_t^T R_t^{-1} B_t \boldsymbol{u}_t + \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1} \right)
\end{aligned}
$$

The current state belief, before the measurement update, is the following distribution:

$$
\Psi(\boldsymbol{x}_t) = \mathcal{C}\left(\boldsymbol{x}_t; \widetilde{P}_t, \widetilde{\boldsymbol{h}}_t, \widetilde{g}_t \right) \tag{2.23}
$$

Simplifications

The equations in Section 2.1.2 are horrific, but they can be reduced to something manageable by applying the Woodbury Matrix Identity (Lemma 1) to the following component:

$$
\left(A_t^T R_t^{-1} A_t + \Sigma_{t-1}^{-1} \right)^{-1} = \left(\Sigma_{t-1} - \Sigma_{t-1} A_t^T \left(R_t + A_t \Sigma_{t-1} A_t^T \right)^{-1} A_t \Sigma_{t-1} \right) \tag{2.24}
$$

Then letting,

$$
\overline{\Sigma}_t = R_t + A_t \Sigma_{t-1} A_t^T \tag{2.25}
$$

Remark 6 *If you don't follow the preceding steps, you will end with a perfectly valid set of equations but they won't be identical to the standard Kalman Filter Algorithm (defined in [8, 9]), which isn't as rewarding.*

Simplifying Equation 2.21

$$
\begin{aligned}
\widetilde{P}_t &= R_t^{-1} - \left(A_t^T R_t^{-1}\right)^T \left(\Sigma_{t-1} - \Sigma_{t-1} A_t^T \left(R_t + A_t \Sigma_{t-1} A_t^T\right)^{-1} A_t \Sigma_{t-1}\right) \left(A_t^T R_t^{-1}\right) \\
&= R_t^{-1} - \left(A_t^T R_t^{-1}\right)^T \left(\Sigma_{t-1} - \Sigma_{t-1} A_t^T \overline{\Sigma}_t^{-1} A_t \Sigma_{t-1}\right) \left(A_t^T R_t^{-1}\right) \\
&= R_t^{-1} - R_t^{-1} \left(A_t \Sigma_{t-1} A_t^T\right) R_t^{-1} + R_t^{-1} \left(A_t \Sigma_{t-1} A_t^T\right) \overline{\Sigma_t^{-1}} \left(A_t \Sigma_{t-1} A_t^T\right) R_t^{-1} \\
&= R_t^{-1} - R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) R_t^{-1} + R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) \overline{\Sigma_t^{-1}} \left(\overline{\Sigma}_t - R_t\right) R_t^{-1} \\
&= R_t^{-1} - R_t^{-1} \overline{\Sigma}_t R_t^{-1} - R_t^{-1} - R_t^{-1} \left(I - R_t \overline{\Sigma}_t^{-1}\right) \left(I - \left(R_t \overline{\Sigma}_t\right)^{-1}\right) \\
&= 2R_t^{-1} - R_t^{-1} \overline{\Sigma}_t R_t^{-1} + R_t^{-1} \left(I - R_t \overline{\Sigma}_t^{-1} - \left(R_t \overline{\Sigma}_t\right)^{-1} + I\right) \\
&= 2R_t^{-1} - R_t^{-1} \overline{\Sigma}_t R_t^{-1} - 2R_t^{-1} + \overline{\Sigma}_t^{-1} + R_t^{-1} \overline{\Sigma}_t R_t^{-1} \\
&= \overline{\Sigma}_t^{-1}
\end{aligned}
\tag{2.26}
$$

Simplifying Equation 2.22:

$$
\begin{aligned}
\widetilde{\boldsymbol{h}}_t &= R_t^{-1} B_t \boldsymbol{u}_t + R_t^{-1} A_t \left(\Sigma_{t-1} - \Sigma_{t-1} A_t^T \left(R_t + A_t \Sigma_{t-1} A_t^T\right)^{-1} A_t \Sigma_{t-1}\right) \left(-A_t^T R_t^{-1} B_t \boldsymbol{u}_t + \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1}\right) \\
&= R_t^{-1} B_t \boldsymbol{u}_t + R_t^{-1} A_t \left(\Sigma_{t-1} - \Sigma_{t-1} A_t^T \overline{\Sigma}_t^{-1} A_t \Sigma_{t-1}\right) \left(-A_t^T R_t^{-1} B_t \boldsymbol{u}_t + \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1}\right) \\
&= R_t^{-1} B_t \boldsymbol{u}_t - R_t^{-1} \left(A_t \Sigma_{t-1} A_t^T\right) R_t^{-1} B_t \boldsymbol{u}_t + R_t^{-1} \left(A_t \Sigma_{t-1} A_t^T\right) \overline{\Sigma}_t^{-1} \left(A_t \Sigma_{t-1} A_t^T\right) R_t^{-1} B_t \boldsymbol{u}_t \\
&\quad + R_t^{-1} A_t \left(\Sigma_{t-1} \Sigma_{t-1}^{-1}\right) \boldsymbol{\mu}_{t-1} - R_t^{-1} \left(A_t \Sigma_{t-1} A_t^T\right) \overline{\Sigma}_t^{-1} A_t \left(\Sigma_t \Sigma_t^{-1}\right) \boldsymbol{\mu}_{t-1} \\
&= R_t^{-1} B_t \boldsymbol{u}_t - R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) R_t^{-1} B_t \boldsymbol{u}_t + R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) \overline{\Sigma}_t^{-1} \left(\overline{\Sigma}_t - R_t\right) R_t^{-1} B_t \boldsymbol{u}_t \\
&\quad + R_t^{-1} A_t \boldsymbol{\mu}_{t-1} - R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) \overline{\Sigma}_t^{-1} A_t \boldsymbol{\mu}_{t-1} \\
&= R_t^{-1} \left(A_t \boldsymbol{\mu}_{t-1} + B_t \boldsymbol{u}_t\right) - R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) \left(I - \overline{\Sigma}_t^{-1} \left(\overline{\Sigma}_t - R_t\right)\right) R_t^{-1} B_t \boldsymbol{u}_t - R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) \overline{\Sigma}_t^{-1} A_t \boldsymbol{\mu}_t \\
&= R_t^{-1} \left(A_t \boldsymbol{\mu}_{t-1} + B_t \boldsymbol{u}_t\right) - R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) \overline{\Sigma}_t^{-1} \left(R_t R_t^{-1}\right) B_t \boldsymbol{u}_t - R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) \overline{\Sigma}_t^{-1} A_t \boldsymbol{\mu}_t \\
&= R_t^{-1} \left(A_t \boldsymbol{\mu}_{t-1} + B_t \boldsymbol{u}_t\right) - R_t^{-1} \left(\overline{\Sigma}_t - R_t\right) \overline{\Sigma}_t^{-1} \left(A_t \boldsymbol{\mu}_{t-1} - B_t \boldsymbol{u}_t\right) \\
&= R_t^{-1} \left(I - \left(\overline{\Sigma}_t - R_t\right) \overline{\Sigma}_t^{-1}\right) \left(A_t \boldsymbol{\mu}_{t-1} + B_t \boldsymbol{u}_t\right) \\
&= \left(R_t^{-1} R_t\right) \overline{\Sigma}_t^{-1} \left(A_t \boldsymbol{\mu}_{t-1} + B_t \boldsymbol{u}_t\right) \\
&= \overline{\Sigma}_t^{-1} \left(A_t \boldsymbol{\mu}_{t-1} + B_t \boldsymbol{u}_t\right)
\end{aligned}
\tag{2.27}
$$

From the definition of the information vector [7]:

$$
\begin{aligned}
\widetilde{\boldsymbol{h}}_t &= \widetilde{P}_t \widetilde{\boldsymbol{\mu}}_t \\
\therefore \widetilde{\boldsymbol{\mu}}_t &= \widetilde{P}_t^{-1} \widetilde{\boldsymbol{h}}_t \\
&= \left(\overline{\Sigma}_t^{-1} \overline{\Sigma}_t\right) \left(A_t \boldsymbol{\mu}_{t-1} + B_t \boldsymbol{u}_t\right) \\
&= A_t \boldsymbol{\mu}_{t-1} + B_t \boldsymbol{u}_t
\end{aligned}
\tag{2.28}
$$

Lemma 1 (Specialised Woodbury Inversion Identity[a]) *For any invertible quadratic matrices R and Q and any matrix P with appropriate dimensions, the following holds true*

$$(R + PQP^T)^{-1} = R^{-1} - R^{-1}P(Q^{-1} + P^T R^{-1} P)^{-1} P^T R^{-1}$$

Proof: *Define $\Psi = (Q^{-1} + P^T R^{-1} P)^{-1}$. It suffices to show that*

$$(R^{-1} - R^{-1}P\Psi P^T R^{-1})(R + PQP) = I$$

This is shown through a series of transformations

$$\begin{aligned}
&= R^{-1}R - R^{-1}PQP^T - R^{-1}P\Psi P^T R^{-1}R + R^{-1}P\Psi P^T R^{-1}PQP^T \\
&= I + R^{-1}PQP^T - R^{-1}P\Psi P^T - R^{-1}P\Psi P^T R^{-1}PQP^T \\
&= I + R^{-1}P\left[QP^T - \Psi P^T - \Psi P^T R^{-1}PQP^T\right] \\
&= I + R^{-1}P\left[QP^T - \Psi Q^{-1}QP^T - \Psi P^T R^{-1}PQP^T\right] \\
&= I + R^{-1}P\left[QP^T - \Psi\left[Q^{-1} + P^T R^{-1}P\right]QP^T\right] \\
&= I + R^{-1}P\left[QP^T - \Psi\Psi^{-1}QP^T\right] \\
&= I + R^{-1}P\left[I - I\right]QP^T \\
&= I
\end{aligned}$$

[a] This is directly stolen, with a few added steps, from [9]. This is also referred to as the Sherman-Morrison-Woodbury Inversion Identity.

2.2 Part 2: Measurement Update

The state estimate is finally completed by incorporating the current measurement, $\boldsymbol{z}_{E,t}$.

$$\delta_{\boldsymbol{z}(t)\to\boldsymbol{x}(t)}(\boldsymbol{x}_t) = \delta_{\phi_{\boldsymbol{z}_t}\to\phi_{\boldsymbol{x}_t}}(\boldsymbol{x}_t)\cdot\Psi(\boldsymbol{x}_t) \tag{2.29}$$

2.2.1 Canonical From Representation

Once again a CPD is rearranged into a joint density:

$$\begin{aligned}
\phi_{\boldsymbol{z}_t}(\boldsymbol{x}_t, \boldsymbol{z}_t) &= \mathcal{N}\left(\boldsymbol{z}_t | C_t\boldsymbol{x}_t, Q_t\right) \\
&= \frac{1}{(2\pi)^{k/2}|Q_t|^{(1/2)}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{z}_t - C_t\boldsymbol{x}_t\right)^T Q_t^{-1}\left(\boldsymbol{z}_t - C_t\boldsymbol{x}_t\right)\right\}
\end{aligned} \tag{2.30}$$

$$\begin{aligned}
&-\frac{1}{2}\left(\boldsymbol{z}_t - C_t\boldsymbol{x}_t\right)^T Q_t^{-1}\left(\boldsymbol{z}_t - C_t\boldsymbol{x}_t\right) \\
&= \begin{bmatrix}\boldsymbol{x}_t^T & \boldsymbol{z}_t^T\end{bmatrix}\begin{bmatrix}C_t^T Q_t^{-1} C_t & -C_t^T Q_t^{-1} \\ -Q_t^{-1}C_t^T & Q_t^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_t \\ \boldsymbol{z}_t\end{bmatrix} \\
&= \left(\boldsymbol{Z}_t\right)^T P_{\boldsymbol{z}_t}\left(\boldsymbol{Z}_t\right)
\end{aligned}$$

Where,

$$\boldsymbol{Z}_t = \begin{bmatrix}\boldsymbol{x}_t \\ \boldsymbol{z}_t\end{bmatrix} \tag{2.31}$$

$$P_{\boldsymbol{z}_t} = \begin{bmatrix}C_t^T Q_t^{-1} C_t & -C_t^T Q_t^{-1} \\ -Q_t^{-1}C_t^T & Q_t^{-1}\end{bmatrix} \tag{2.32}$$

Resulting in the zero mean distribution,

$$\phi_{\boldsymbol{z}_t}(\boldsymbol{x}_t, \boldsymbol{z}_t) = \mathcal{C}\left(\boldsymbol{x}_t, \boldsymbol{z}_t; P_{\boldsymbol{z}_t}, \boldsymbol{0}, g_{\boldsymbol{z}_t}\right) \tag{2.33}$$

Where,

$$g_{\boldsymbol{z}_t} = -\ln\left\{(2\pi)^{k/2}|Q_t|^{1/2}\right\} \tag{2.34}$$

Observations

The measurement of the current state is made, $\boldsymbol{z}_t = \boldsymbol{z}_{E,t}$. According to [7]:

$$\widetilde{P}_{\boldsymbol{z}_t} = C_t^T Q_t^{-1} C_t \tag{2.35}$$

$$\widetilde{\boldsymbol{h}}_{\boldsymbol{z}_t} = C_t^T Q_t^{-1} \boldsymbol{z}_{E,t} \tag{2.36}$$

$$\widetilde{g}_{\boldsymbol{z}_t} = -\ln\left\{(2\pi)^{k/2}|Q_t|^{1/2}\right\} - \frac{1}{2}\boldsymbol{z}_t^T Q_t^{-1} \boldsymbol{z}_t \tag{2.37}$$

Therefore,

$$\delta_{\boldsymbol{z}(t)\to\boldsymbol{x}(t)}(\boldsymbol{x}_t) = \phi_{\boldsymbol{z}(t)}(\boldsymbol{x}_t) = \mathcal{C}\left(\boldsymbol{x}_t; \overline{P}_{\boldsymbol{z}_t}, \overline{\boldsymbol{h}}_{\boldsymbol{z}_t}, \overline{g}_{\boldsymbol{z}_t}\right) \tag{2.38}$$

2.2.2 Update and Simplifications

Finally, the measurement update is performed:

$$\begin{aligned}\delta_{\phi_{\boldsymbol{x}_t}\to\phi_{\boldsymbol{x}_{t+1}}}(\boldsymbol{x}_t) &= \delta_{\phi_{\boldsymbol{z}_t}\to\phi_{\boldsymbol{x}_t}}(\boldsymbol{x}_t)\cdot\Psi(\boldsymbol{x}_t)\\ &= \mathcal{C}\left(\boldsymbol{x}_t; \widetilde{P}_{\boldsymbol{z}_t}, \widetilde{\boldsymbol{h}}_{\boldsymbol{z}_t}, \widetilde{g}_{\boldsymbol{z}_t}\right)\cdot\mathcal{C}\left(\boldsymbol{X}_t; \widetilde{P}_t, \widetilde{\boldsymbol{h}}_t, \widetilde{g}_t\right)\\ &= \mathcal{C}\left(\boldsymbol{x}_t; \widetilde{P}_t + \widetilde{P}_{\boldsymbol{z}_t}, \widetilde{\boldsymbol{h}}_t + \widetilde{\boldsymbol{h}}_{\boldsymbol{z}_t}, \widetilde{g}_t + \widetilde{g}_{\boldsymbol{z}_t}\right)\\ &= \mathcal{C}\left(\boldsymbol{x}_t; \overline{P}_t, \overline{\boldsymbol{h}}_t, \overline{g}_t\right)\end{aligned} \tag{2.39}$$

Where,

$$\overline{P}_t = \overline{\Sigma}_t^{-1} + C_t^T Q_t^{-1} C_t \tag{2.40}$$

$$\overline{\boldsymbol{h}}_t = C_t^T Q_t^{-1} \boldsymbol{z}_{E,t} + \overline{\Sigma}_t^{-1}\overline{\boldsymbol{\mu}}_t \tag{2.41}$$

The precision matrix, $\overline{P}_t$, is defined as the inverse of the covariance matrix, therefore:

$$\Sigma_t = \left(\overline{\Sigma}_t^{-1} + C_t^T Q_t^{-1} C_t\right)^{-1} \tag{2.42}$$

Using Lemma 1 again,

$$\Sigma_t = \left(\overline{\Sigma}_t - \overline{\Sigma}_t C_t^T \left(Q_t + C_t\overline{\Sigma}_t C_t^T\right)^{-1} C_t\overline{\Sigma}_t\right) \tag{2.43}$$

Defining the Kalman Gain, K_t, as:

$$K_t = \overline{\Sigma}_t C_t^T \left(Q_t + C_t\overline{\Sigma}_t C_t^T\right)^{-1} \tag{2.44}$$

$$\therefore \Sigma_t = (I - K_t C_t)\,\overline{\Sigma}_t \tag{2.45}$$

From the definition of the information vector [7]:

$$\begin{aligned}
&\overline{\boldsymbol{h}}_t = \Sigma_t^{-1}\boldsymbol{\mu}_t \\
&\therefore \boldsymbol{\mu}_t = \Sigma_t\overline{\boldsymbol{h}}_t \\
&= (I - K_tC_t)\,\overline{\Sigma}_t\left(C_t^TQ_t^{-1}\boldsymbol{z}_{E,t} + \overline{\Sigma}_t^{-1}\widetilde{\boldsymbol{\mu}}_t\right) \\
&= (I - K_tC_t)\,\overline{\Sigma}_tC_t^TQ_t^{-1}\boldsymbol{z}_{E,t} + (I - K_tC_t)\left(\overline{\Sigma}_t\overline{\Sigma}_t^{-1}\right)\widetilde{\boldsymbol{\mu}}_t \\
&= \left(\overline{\Sigma}_tC_t^T - K_tC_t\overline{\Sigma}_tC_t^T\right)Q_t^{-1}\boldsymbol{z}_{E,t} + (I - K_tC_t)\,\widetilde{\boldsymbol{\mu}}_t
\end{aligned} \tag{2.46}$$

A quick sidestep,

$$\begin{aligned}
&K_t = \overline{\Sigma}_tC_t^T\left(Q_t + C_t\overline{\Sigma}_tC_t^T\right)^{-1} \\
&K_t\left(Q_t + C_t\overline{\Sigma}_tC_t^T\right) = \overline{\Sigma}_tC_t^T \\
&K_tQ_t = \overline{\Sigma}_tC_t^T - K_tC_t\overline{\Sigma}_tC_t^T
\end{aligned} \tag{2.47}$$

Now,

$$\begin{aligned}
\boldsymbol{\mu}_t &= (K_tQ_t)\,Q_t^{-1}\boldsymbol{z}_{E,t} + (I - K_tC_t)\,\widetilde{\boldsymbol{\mu}}_t \\
&= K_t\left(Q_tQ_t^{-1}\right)\boldsymbol{z}_{E,t} + \overline{\boldsymbol{\mu}}_t - K_tC_t\widetilde{\boldsymbol{\mu}}_t \\
&= \widetilde{\boldsymbol{\mu}}_t + K_t\left(\widetilde{\boldsymbol{\mu}}_t - C_t\boldsymbol{z}_{E,t}\right)
\end{aligned} \tag{2.48}$$

Remark 7 *You may notice I have neglected the scalar $\overline{g}_t$, it didn't really contain any of the really interesting equations. I assume it evaluates to a valid normalization constant as we have operated in a linear Gaussian system.*

References

[1] D. Barber. 'Belief Networks' in *Bayesian Reasoning and Machine Learning*, 1^{st} ed., vol. 1. Cambridge, U.K: Cambridge University Press, 2012, pp. 29-58.

[2] D. Barber. 'Graphical Models' in *Bayesian Reasoning and Machine Learning*, 1^{st} ed., vol. 1. Cambridge, U.K: Cambridge University Press, 2012, pp. 58-77.

[3] D. Barber. 'The Junction Tree Algorithm' in *Bayesian Reasoning and Machine Learning*, 1^{st} ed., vol. 1. Cambridge, U.K: Cambridge University Press, 2012, pp. 102-127.

[4] J. Du Preez. 'Inference: The Junction-Tree Algorithm'. Internet: http://research.ee.sun.ac.za/pgms/instructions/week5_junctiontree.pdf, 2016 [Apr. 26, 2016].

[5] D. Koller and N. Friedman. 'The Bayesian Network Representation' in *Probabilistic Graphical Models: Principles and Techniques*, 1^{st} ed., vol. 1. Cambridge, Massachusetts: The MIT Press, 2009, pp. 45-83.

[6] D. Koller and N. Friedman. 'Exact Inference' in *Probabilistic Graphical Models: Principles and Techniques*, 1^{st} ed., vol. 1. Cambridge, Massachusetts: The MIT Press, 2009, pp. 345-376.

[7] D. Koller and N. Friedman. 'Inference in Hybrid Networks' in *Probabilistic Graphical Models: Principles and Techniques*, 1^{st} ed., vol. 1. Cambridge, Massachusetts: The MIT Press, 2009, pp. 609-611.

[8] S. Thrun et al. 'Recursive State Estimation' in *Probabilistic Robotics*, 1^{st} ed., vol. 1. Cambridge, Massachusetts: The MIT Press, 2009, pp. 14-35.

[9] S. Thrun et al. 'Gaussian Filters' in *Probabilistic Robotics*, 1^{st} ed., vol. 1. Cambridge, Massachusetts: The MIT Press, 2009, pp. 39-53.