Spectral Methods for Learning Kalman Filters

Instructor: Sham Kakade

1 Kalman Filters

We now summarize a simplified version of linear Gaussian time series. Here, we assume that the transition noise and observation noise are stationary.

Assume that:

$$h_{t+1} = Th_t + \eta$$

where η is a multivariate normal (with some fixed unknown covariance matrix). Also, assume:

$$x_t = Oh_t + \varepsilon$$

where ε is multivariate normal (with some fixed unknown covariance matrix). To completely specify the model, we must specify the distribution under which h_1 is drawn from.

1.1 Stationary Kalman filters

Let us assume that T, O, and both noise covariance matrices are full rank. One can show that the posterior distribution of $Pr(h_t|x_1, \ldots x_{t-1})$ will converge to a multivariate normal, with some asymptotic covariance distribution. Let us say this distribution is $N(h_\infty, \Sigma_\infty)$.

For simplicity, let us assume that the initial hidden state is sampled from this distribution, i.e. $h_1 \sim N(h_\infty, \Sigma_\infty)$.

We are interested in keeping track of the hidden state and predicting the next observation. Let us define:

$$g_t = \mathbb{E}[h_t|x_{<t}]$$

These are the quantities that we would like to compute.

The Kalman filter says that these expressions have the following form. Initially,

$$g_1 = h_\infty$$

and for all future times:

$$\begin{aligned} g_{t+1} &= Tg_t + K(x_t - Og_t) \\ \mathbb{E}[x_{t+1}|x_{<t+1}] &= Og_{t+1} \end{aligned}$$

Here K is the Kalman gain matrix, and $x_t - Og_t$ is often referred to as the "innovation", "measurement residual", or "measurement error". The KF takes this particularly simple form as we have assumed that h_1 is sampled from the asymptotic distribution and that our noise and transition model are stationary. Otherwise, K would vary with time.

Note that these are simple matrix update rules.

1.2 Agnostic Assumptions and best fit Kalman Filters

The more general class of Gaussian linear models is where:

$$h_{t+1} = Th_t + \eta_t \quad \text{and} \quad x_t = Oh_t + \varepsilon_t$$

where both noise terms are time dependent Gaussian noise. Again, if these noise covariances are known, then the Kalman filter is a simple way to compute conditional expectations (and posterior distributions). Here, the Kalman gain matrix K will be time dependent.

It is straightforward to see in this more general setting that conditional expectation $\mathbb{E}[x_t|x_{<t}]$ is linear in $x_{<t}$. In fact, one can view the Kalman filter as a concise way of computing this conditional expectation (which exploits the time series structure).

Now among the more general class of state-space models that we are considering, we can ask the question of what the best linear prediction of $\mathbb{E}[x_t|x_{<t}]$ is? By linear, we mean in terms of $x_{<t}$.

Lemma 1.1. *For any state space model, where:*

$$\mathbb{E}[h_{t+1}|h_t] = Th_t \quad \textit{and} \quad \mathbb{E}[x_t|h_t] = Oh_t$$

Let the best linear prediction of $\mathbb{E}[x_t|x_{<t}]$ be $w \cdot x_{<t}$. There exists a Gaussian noise model (with T and O the same but with appropriately chosen time varying covariance matrices), such that the Kalman filters computation of $\mathbb{E}[x_t|x_{<t}]$ is identical to $w \cdot x_{<t}$.

For example, even if the model is an HMM, the best linear prediction (as a function of the entire history) can be computed by a Kalman filter (with appropriately chosen noise). We can view this lemma as showing how the best fit Gaussian noise model/Kalman filters are "robust" even when the underlying dynamics are non-Gaussian.

2 In Our Transformed Representation

Assumption 1 (Stationarity and Full Rank)**.** *Assume that:*

- *T and O are full rank.*
- *The model has stationary Gaussian noise (with full rank covariance matrices).*
- *h_1 is a multivariate normal (with the asymptotic mean and covariance matrix). This implies the Kalman gain matrix is stationary.*

Recall our transformed representation:

$$\widetilde{h}_t = Mh_t \quad \text{and} \quad \widetilde{T} = MTM^{-1}$$

where $h_t = M^{-1}\widetilde{h}_t$ (since M is invertible) and

$$\mathbb{E}[\widetilde{h}_{t+1}|\widetilde{h}_t] = \widetilde{T}\widetilde{h}_t \quad \text{and} \quad \mathbb{E}[x_t|\widetilde{h}_t] = U\widetilde{h}_t$$

Also, recall that we can recover both $\widetilde{T}$ and U.

Define:

$$\widetilde{g}_t = \mathbb{E}[\widetilde{h}_t|x_{<t}] = Mg_t$$

Lemma 2.1. *In this representation, the KF is:*

$$\begin{aligned}
\widetilde{g}_1 &= Mg_1 = Mh_1 \\
\widetilde{g}_{t+1} &= \widetilde{T}\widetilde{g}_t + \widetilde{K}(x_t - U\widetilde{g}_t) \\
\mathbb{E}[x_t|x_{<t}] &= U\widetilde{g}_t
\end{aligned}$$

where $\widetilde{K} = MK$.

Proof. First, note that:

$$\mathbb{E}[x_t|x_{<t}] = \mathbb{E}[\mathbb{E}[x_t|\widetilde{h}_t]|x_{<t}] = \mathbb{E}[U\widetilde{h}_t|x_{<t}] = U\widetilde{g}_t$$

From the original KF, we have

$$g_{t+1} = Tg_t + K(x_t - Og_t)$$

By multiplying by M, we have:

$$\begin{aligned} \widetilde{g}_{t+1} &= MTg_t + MK(x_t - Og_t) \\ &= MTM^{-1}\widetilde{g}_t + \widetilde{K}(x_t - \mathbb{E}[x_t|x_{<t}]) \\ &= \widetilde{T}\widetilde{g}_t + \widetilde{K}(x_t - U\widetilde{g}_t) \end{aligned}$$

which completes the proof. □

3 Learning the KF and "bottleneck prediction"

As we have $\widetilde{T}$ and U already, all that remains to specify is $\widetilde{g}_1$ and $\tilde{K}$.

Theorem 3.1. *Assume our Stationarity and Full Rank assumption. Let the "thin" SVD of the cross correlation matrix at some timestep* 1 *be* $E[x_2x_1^\top] = UDV^\top$*. Then we have that* $M = U^\top O$ *is invertible. Define*

$$\Sigma_{11} = \mathbb{E}[(x_1 - \mathbb{E}[x_1])^\top(x_1 - E[x_1])^\top] \quad \textit{and} \quad \Sigma_{21} = \mathbb{E}[(x_2 - \mathbb{E}[x_2])(x_1 - \mathbb{E}[x_1])^\top]$$

Then our Kalman filter uses the following parameters:

$$\begin{aligned} \widetilde{T} &= (U^\top\mathbb{E}[x_3x_1^\top])(U^\top\mathbb{E}[x_2x_1^\top])^+ \\ g_1 &= U^\top E[x_1] \\ \widetilde{K} &= U^\top\Sigma_{21}\Sigma_{11}^{-1} \end{aligned}$$

where the inverse exists.

Proof. By our previous lemma, we have that:

$$\begin{aligned} \mathbb{E}[x_2|x_1] &= U\widetilde{g}_2 \\ &= U\widetilde{T}\widetilde{g}_1 + U\widetilde{K}(x_1 - U\widetilde{g}_1) \\ &= \mathbb{E}[x_2] + U\widetilde{K}(x_1 - E[x_1]) \end{aligned}$$

i.e.

$$\mathbb{E}[x_2 - \mathbb{E}[x_2]|x_1] = U\widetilde{K}(x_1 - E[x_1])$$

Multiplying by $(x_1 - E[x_1])^\top$ and taking expectations:

$$\Sigma_{21} = U\widetilde{K}\Sigma_{11}$$

Now, we have have that:

$$U\widetilde{K} = \Sigma_{21}\Sigma_{11}^{-1}$$

Since $U^\top U = \mathrm{I}$ (as U has orthonormal columns), we have our result. □