The 1d Kalman Filter

Richard Turner

This is a Jekyll and Hyde of a document and should really be split up. We start with Jekyll which contains a very short derivation for the 1d Kalman filter, the purpose of which is to give intuitions about its more complex cousin. I find the Kalman filter / linear Gaussian state space model thing tough to intuit. This is an attempt to remedy this. Hyde can be found afterwards - here I show the connection between linear Gaussian state space models to Gaussian processes and the Wiener filter. This document is missing some pictures and references, apologies for that.

1 Understanding the forward model

Let’s define the probabilistic model, together with some ways of interpreting it, and recap the task the Kalman filter solves for us. The 1d linear Gaussian state space model can be written:

\[
P(x_1) = \text{Norm}(\mu, \sigma^2_1) \tag{1}
\]

\[
P(x_t|x_{t-1}) = \text{Norm}(\lambda x_{t-1}, \sigma^2) \tag{2}
\]

\[
P(y_t|x_t) = \text{Norm}(x_t, \sigma^2_y) \tag{3}
\]

Notice we are free to rescale the parameters, and therefore the latents, and so the restriction of the emission weights to unity is unimportant.

There are several ways of thinking about this model and the most useful will depend on the situation. Let’s think about generating data according to the forward model. At each time step observations are produced like they are in a factor analysis model. However, rather than having uncorrelated latents at each time step drawn from a white Gaussian, the latents are now correlated across time. For instance if \( \lambda \to 1 \) and \( \sigma^2 \to 0 \) the latents we generate become very smooth and slowly varying, and in the other limit \( \lambda \to 0 \) (and \( \sigma^2 \to 1 \)) the latents are highly variable and we recover factor analysis. This model, then, specifies a more general prior distribution over functions for the latents than factor analysis, but otherwise is identical. For the keen, later on I give another more general interpretation for the LGSSM model, but this is really all you need to keep in your head.

Let’s give a concrete example where this might be a useful model. Imagine you are on a geography field trip trying to map out the depth of a lake along its length. The lake is cold, deep, and murky so you can’t see the bottom, nor would you want to swim in it, but you do have a boat. You decide to measure the depth using a long piece of rope with a rock tied to
the end. You lower the rope into the water until you feel it hit the bottom, and measure the length of rope to get an estimate of the depth. Being an organised person, you move equal distances along the lake between each measurement. However, due to currents in the water, an inability to determine precisely when the rock hits the bottom, and your boat drifting a bit, the depth measurements are noisy. You estimate that they might fluctuate by as much as half a metre from the true depth. However, your geography teacher tells you that the lakes in this area are known to have smoothly undulating bottoms (\textit{?}). By injecting this prior information - that the depth is unlikely to vary much between two nearby points - you could estimate the depth at any point with an accuracy greater than that intrinsic to the rope and rock. \textit{Insert some pictures of sampling from the forward model here}

2 Inference

Given some data \( y_{1:T} \) and a setting of the parameters \( \mu, \sigma^2_1, \) and \( \sigma^2_y \), there are several inferences we may wish to make. We may wish to infer the posterior distribution over one of the latents online: \( P(x_t | y_{1:t}) \). This is exactly the problem the Kalman Filter solves. There are other interesting distributions, which it does not return. For instance \( P(x_t | y_{1:T}) \) (for this we need the Kalman smoother), \( P(x_{1:T} | y_{1:T}) \) (for this the Wiener filter), and \( P(x_{t+1} | y_{1:t}) \) (which crops up any time we want to make predictions eg. GPs).

Given the structure of the graph, it seems intuitive that we should be able to use the quantity \( P(x_{t-1} | y_{1:t-1}) \) to help us calculate \( P(x_t | y_{1:t}) \). If this were true, it would enable us to calculate the distribution of \( x_t \) recursively and therefore efficiently. Let’s see if we can manipulate \( P(x_t | y_{1:t}) \) to do this. The general approach is to introduce the latent \( x_{t-1} \), and use the (in)dependencies specified by the graphical model to simplify the result.

\[
P(x_t | y_{1:t}) = \int P(x_t, x_{t-1} | y_{1:t}) dx_{t-1}
\]

\[
= \frac{1}{P(y_{1:t})} \int P(x_t, x_{t-1}, y_{1:t}) dx_{t-1}
\]

\[
= \frac{P(y_{1:t-1})}{P(y_{1:t})} \int P(x_t | x_{t-1}) P(y_t | x_t) P(x_{t-1} | y_{1:t-1}) dx_{t-1}
\]

\[
= \frac{1}{Z} P(y_t | x_t) \int P(x_t | x_{t-1}) P(x_{t-1} | y_{1:t-1}) dx_{t-1}
\]

\[
= \frac{1}{Z} P(y_t | x_t) P(x_t | y_{1:t-1})
\]

\[
= \frac{1}{Z} \text{likelihood} \times \text{prior}
\]

The distribution over the current latent is found by combining what the new observation told us about it with what the previous observations told
us. Both the likelihood function and the prior are Gaussian balls in latent space, and so too is the distribution over \( x_t \).

*Insert pictures of inference using the likelihood and the prior distributions here*

Substituting in for these distributions:

\[
P(x_t | y_{1:t}) = \frac{1}{Z} \int \exp \left[ -\frac{1}{2\sigma_t^2} (x_t - \lambda x_{t-1})^2 - \frac{1}{2\sigma_y^2} (y_t - x_t)^2 ight] dx_{t-1}
\]

\[
= \text{Norm}(\langle x_t \rangle, \sigma_t^2)
\]

The integral is performed by completing the square in the exponent for the \( x_{t-1} \) terms. Keeping the terms in \( x_t \) only:

\[
P(x_t | y_{1:t}) = \frac{1}{Z} \exp \left[ -\frac{1}{2} \left( x_t^2 \left( \frac{1}{\sigma_y^2} + \frac{1}{\sigma^2 + \lambda^2\sigma_t^{-1}} \right) \right) 
\]

\[
-2x_t \left( \frac{\lambda \langle x_{t-1} \rangle}{\sigma^2 + \lambda^2\sigma_t^{-1}} + \frac{y_t}{\sigma_y^2} \right) \]

From which the means and variances can be recovered by moment matching, giving the Kalman filter recursions:

\[
\langle x_t \rangle = \lambda \langle x_{t-1} \rangle + K_t (y_t - \lambda \langle x_{t-1} \rangle) 
\]

\[
\sigma_t^2 = \sigma_y^2 K_t 
\]

\[
K_t = \frac{\sigma^2 + \lambda^2\sigma_t^{-1}}{\sigma^2 + \sigma_y^2 + \lambda^2\sigma_t^{-1}} 
\]

These are the celebrated Kalman filter recursions for our simple 1d model and \( K_t \) is known as the Kalman gain. We can understand them better by considering the form of the prior distribution. It is exactly as we might have expected: \( P(x_t | y_{1:t-1}) = \text{Norm}(\lambda \langle x_{t-1} \rangle, \lambda^2\sigma_t^{-1} + \sigma^2) \). (Nb. If our prior beliefs about \( x \) are that they do not blow up exponentially we must have \( \lambda \leq 1 \).) We interpret the Kalman filter recursions as follows:

\[
\langle x_t \rangle = \text{prior estimate for } x_t + \text{correction} 
\]

\[
\text{correction} = K_t \times (\text{ML estimate for } x_t - \text{prior estimate for } x_t) 
\]

\[
\sigma_t^2 = \text{ML variance of } x_t \times K_t 
\]

\[
K_t = \frac{\text{prior variance of } x_t}{\text{ML variance of } x_t + \text{prior variance of } x_t} 
\]

You can check that this does all the right things as the ML variance of \( x_t \) tends to infinity or zero.
3 Connection to GPs

Here’s the second way of thinking about the model that I promised earlier. It makes the connection with GPs and the Wiener filter. (If you are reading this for a simple introduction: you can ignore this section without missing anything.) Earlier we made an analogy to FA where the generation of data \((y_t)\) at each time step mapped to a single realisation from the FA model. There is an alternative: We could map a draw from the whole chain \((y_{1:T})\) to a single realisation from the FA model. Different draws from the entire chain are equivalent to different realisations from the FA model. Using the example above, FA would specify a white Gaussian prior over the latents: \(P(x_{1:T}) = I\). The linear dynamical state space model specifies a correlated prior parameterised by \(\lambda\) and \(\sigma^2\). Let’s be a little more explicit about this correspondence by finding the precise prior distribution the LDSSM specifies.

\[
P(x_{1:T}) = P(x_1) \prod_{t=2}^T P(x_t|x_{t-1}) \quad (21)
\]

\[
= \frac{1}{Z} \exp \left[ -\frac{1}{2} \left( \frac{x_1^2}{\sigma_1^2} + \sum_{t=2}^T \frac{1}{\sigma^2} (x_t - \lambda x_{t-1})^2 \right) \right] \quad (22)
\]

\[
= \frac{1}{Z} \exp \left[ -\frac{1}{2} \left( \frac{x_1^2}{\sigma_1^2} + \frac{\lambda^2}{\sigma_2^2} \right) - 2x_1x_2 \frac{\lambda}{\sigma_2^2} 
+ x_2^2 \frac{1 + \lambda^2}{\sigma_2^2} + 2x_3x_2 \frac{\lambda}{\sigma_2^2} + x_3^2 \frac{1 + \lambda^2}{\sigma_2^2} + 2x_4x_3 \frac{\lambda}{\sigma_2^2} + ... \right] \quad (23)
\]

\[
= \text{Norm}(0, \Sigma) \quad (25)
\]

Where the last line follows from the fact that a linear transformation of Gaussian random variables is also Gaussian. This makes the connection to FA explicit where \(\Sigma = I\). In this sense the LGSSM is more general than FA. A quick caveat should be added with regard to the emission weights: the LGSSM specifies causal emissions (the emission at time \(t\) depends only on the latents at time \(t\)), FA does not. For this reason, this model is more restrictive than FA (there are lots of zeros in the weight matrix \(W\) in \(P(y_{1:T}|x_{1:T}) = \text{Norm}(W x_{1:T}, \Phi)\)). By matching terms we find:

\[
\Sigma^{-1} = \begin{bmatrix}
\frac{1}{\sigma_1^2} + \frac{\lambda^2}{\sigma^2} & -\frac{\lambda}{\sigma^2} & 0 & 0 & ... \\
-\frac{\lambda}{\sigma^2} & \frac{1+\lambda^2}{\sigma^2} & -\frac{\lambda}{\sigma^2} & 0 & ... \\
0 & -\frac{\lambda}{\sigma^2} & \frac{1+\lambda^2}{\sigma^2} & -\frac{\lambda}{\sigma^2} & ... \\
0 & 0 & -\frac{\lambda}{\sigma^2} & \frac{1+\lambda^2}{\sigma^2} & ... \\
... & ... & ... & ... & ... \\
\end{bmatrix} \quad (26)
\]
If $\sigma^2_1 = \sigma^2_2$ this means: $\Sigma_{i,j} = \frac{\sigma^2}{1-\lambda^2} |i-j|$. The covariance of the prior is stationary (as it must have been due to the shift-invariant specification) and decays exponentially. The power spectrum (eigenvalues) of the covariance matrix are therefore $\propto \frac{1}{1+\omega^2 \log |\lambda|^2}$ so high frequency components decay away quadratically with a characteristic length $\log \lambda$, matching the intuition that values of $\lambda$ near 1 correspond to smoother priors. The LGSSM therefore specifies a stationary Gaussian process prior over functions for the discretely sampled latents. The Wiener filter is the method for returning the full posterior distribution over the latents $P(x_{1:T}|y_{1:T})$ for a model with a Gaussian process prior over the latents and Gaussian emissions. There’s a document about it on my website too.

We have moved from Kalman filters to Gaussian processes, but we can move the other way too. Kalman filters specify their prior using conditional distributions, whereas GPs typically specify the full distribution:

$$P(x_{1:T}|\theta_{GP}) = \mathcal{N}(0, C_{1:T}) (27)$$

$$P(x_t|x_{t-\tau:t-1}, \theta_{SSM}) = \mathcal{N}(Ax_{t-\tau:t-1}, \sigma^2) (28)$$

Firstly we should ensure the GP has the same conditional independence relations (directed graph) as the SSM. Thus any pair of variables $x_i, x_j$ should be uncorrelated if $|i-j| > \tau$ and then $C_{i,j}^{-1} = 0$. Furthermore, as the SSM specifies a conditional which is shift invariant, $C$ will be Toeplitz (and symmetric).

Now we find the conditions on the parameters for equivalence by forming the conditional for the GP (which is a special case of the predictive distribution).

$$P(x_t|x_{t-\tau:t-1}, \theta_{GP}) = \frac{P(x_{t-\tau:t}|\theta_{GP})}{P(x_{t-\tau:t-1}|\theta_{GP})} (29)$$

where

$$P(x_{t_1:t_2}|\theta_{GP}) = \int P(x_{1:T}|\theta_{GP})dx_{1:t_1-1}dx_{t_2+1:T} (30)$$

But from the definition of the GP (consistent marginalisation properties), if:

$$C_{1:T} = \begin{pmatrix} C_{1:t_1-1} & L & M \\ L^T & C_{t_1:t_2} & N \\ M^T & N^T & C_{t_2+1:T} \end{pmatrix} (31)$$
then,
\[ P(x_{t_1} : x_{t_2} | \theta_{GP}) = G_{x_{t_1}:x_{t_2}} (\mathbf{0}, C_{t_1:t_2}) \]  
(32)

and thus;
\[ P(x_{t-\tau:t} | \theta_{GP}) = G_{x_{t-\tau:t}} (\mathbf{0}, C_{t-\tau:t}) \]  
(33)

\[ P(x_{t-\tau:t-1} | \theta_{GP}) = G_{x_{t-\tau:t-1}} (\mathbf{0}, C_{t-\tau:t-1}) \]  
(34)

where the two covariance matrices are related by:
\[ C_{t-\tau:t} = \begin{pmatrix} C_{t-\tau:t-1} & \mathbf{c} \\ \mathbf{c}^T & c \end{pmatrix} \]  
(35)

Completing the square we find (after von Mises, 1964);
\[ P(x_t | x_{t-\tau:t-1}, \theta_{GP}) = G_{x_t} \left( \mathbf{c}^T C_{t-\tau:t-1}^{-1}, c - \mathbf{c}^T C_{t-\tau:t-1}^{-1} \mathbf{c} \right) \]  
(36)

Therefore the conditions for equivalence are:
\[ \Lambda = \mathbf{c}^T C_{t-\tau:t-1}^{-1} \quad \forall t \]  
(37)
\[ \sigma^2 = c - \mathbf{c}^T C_{t-\tau:t-1}^{-1} \mathbf{c} \quad \forall t \]  
(38)

References