K State space filters

K.1 Terminology

There are three helpful terms here:

1. A *predictor* gives a current state from past observations.
2. A *filter* gives a current state from past and current observations.
3. A *smoother* gives a current state from past, current and future observations.

K.2 The Kalman filter

Say we have a sequence of (unknown) states, $\{\lambda_1, \ldots, \lambda_S\}$. Each state depends upon the previous state. In turn, we can only observe the states through another noisy channel, giving observations $\{x_1, \ldots, x_S\}$. We want to say something about the state sequence given the observations. The filter is given by

$$p\left(\lambda_i \mid x_1, \ldots, x_i\right) = \frac{p\left(x_i \mid \lambda_i, x_1, \ldots, x_{i-1}\right) p\left(\lambda_i \mid x_1, \ldots, x_{i-1}\right)}{p\left(x_i \mid x_1, \ldots, x_{i-1}\right)}, \tag{117}$$

$$= \frac{p\left(x_i \mid \lambda_i\right) p\left(\lambda_i \mid x_1, \ldots, x_{i-1}\right)}{\int d\lambda_i' \, p\left(x_i \mid \lambda_i'\right) p\left(\lambda_i' \mid x_1, \ldots, x_{i-1}\right)}, \tag{118}$$

where the integrals are definite over the range of the variable. Note that last term in the numerator is a predictor, it gives a current state from past observations. The predictor is given by

$$p\left(\lambda_i \mid x_1, \ldots, x_{i-1}\right) = \int d\lambda_{i-1} \, p\left(\lambda_i \mid \lambda_{i-1}, x_1, \ldots, x_{i-1}\right) p\left(\lambda_{i-1} \mid x_1, \ldots, x_{i-1}\right), \tag{119}$$

$$= \int d\lambda_{i-1} \, p\left(\lambda_i \mid \lambda_{i-1}\right) p\left(\lambda_{i-1} \mid x_1, \ldots, x_{i-1}\right). \tag{120}$$

The last term is a filter, so it's recursive, and eventually that final term becomes unconditional (i.e., a prior).

Say we model all the distributions as Gaussian.

$$p\left(\lambda_1\right) = \mathcal{N}\left(\lambda_1; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\lambda_1 - \mu\right)^2}{2\sigma^2}\right), \tag{121}$$

$$p\left(\lambda_i \mid \lambda_{i-1}\right) = \mathcal{N}\left(\lambda_i; \lambda_{i-1}, \sigma_\lambda^2\right) = \frac{1}{\sqrt{2\pi}\sigma_\lambda} \exp\left(-\frac{\left(\lambda_i - \lambda_{i-1}\right)^2}{2\sigma_\lambda^2}\right), \tag{122}$$

$$p\left(x_i \mid \lambda_i\right) = \mathcal{N}\left(x_i; \lambda_i, \sigma_x^2\right) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{\left(x_i - \lambda_i\right)^2}{2\sigma_x^2}\right). \tag{123}$$

So the first filter is

$$p\left(\lambda_1 \mid x_1\right) = \frac{p\left(x_1 \mid \lambda_1\right) p\left(\lambda_1\right)}{\int d\lambda_1' \, p\left(x_1 \mid \lambda_1'\right) p\left(\lambda_1'\right)}, \tag{124}$$

$$= \mathcal{N}\left(\lambda_1; \frac{x_1\sigma^2 + \mu\sigma_x^2}{\sigma^2 + \sigma_x^2}, \frac{\sigma^2\sigma_x^2}{\sigma^2 + \sigma_x^2}\right), \tag{125}$$

$$= \mathcal{N}\left(\lambda_1; M_1^+, V_1^+\right), \tag{126}$$

where that second line is a pretty standard result after some messy algebra. Then the first predictor is

$$p\left(\lambda_2 \mid x_1\right) = \int_{-\infty}^{\infty} d\lambda_1 \, p\left(\lambda_2 \mid \lambda_1\right) p\left(\lambda_1 \mid x_1\right), \tag{127}$$

$$= \mathcal{N}\left(\lambda_2; M_1^+, \sigma_\lambda^2 + V_1^+\right), \tag{128}$$

$$= \mathcal{N}\left(\lambda_2; M_1^+, P_2\right), \tag{129}$$

being a convolution of two Gaussians; the variances add. That predictor then replaces the prior for the second iteration:

$$p(\lambda_2 \mid x_1, x_2) = \frac{p(x_2 \mid \lambda_2, x_1)\, p(\lambda_2 \mid x_1)}{\int d\lambda_2'\, p(x_2 \mid \lambda_2', x_1)\, p(\lambda_2' \mid x_1)}, \tag{130}$$

$$= \mathcal{N}\left(\lambda_2; \frac{x_2 P_2 + M_1^+ \sigma_x^2}{P_2 + \sigma_x^2}, \frac{P_2 \sigma_x^2}{P_2 + \sigma_x^2}\right), \tag{131}$$

$$= \mathcal{N}\left(\lambda_2; M_2^+, V_2^+\right). \tag{132}$$

At each iteration, the means $\{M_1^+, \ldots, M_S^+\}$ are the MAP estimates of the states.

K.3 The Kalman smoother

Once you get to the end of a sequence and have an estimate of $\hat{\lambda}_S$, it's possible to say something more about the previous estimates. In particular,

$$p(\lambda_{S-1} \mid x_1, \ldots, x_S) = \int d\lambda_S\, p(\lambda_{S-1} \mid \lambda_S, x_1, \ldots, x_S)\, p(\lambda_S \mid x_1, \ldots, x_S), \tag{133}$$

$$= \int d\lambda_S\, p(\lambda_S \mid x_1, \ldots, x_S)\, p(\lambda_{S-1} \mid \lambda_S, x_1, \ldots, x_{S-1}). \tag{134}$$

Now, the first term is

$$p(\lambda_S \mid x_1, \ldots, x_S) = \mathcal{N}\left(\lambda_S; M_S^+, V_S^+\right) \tag{135}$$

$$= \mathcal{N}\left(\lambda_S; M_S^-, V_S^-\right) \tag{136}$$

and the fractional term evaluates to

$$p(\lambda_{S-1} \mid \lambda_S, x_1, \ldots, x_{S-1}) = \frac{p(\lambda_S \mid \lambda_{S-1}, x_1, \ldots, x_{S-1})\, p(\lambda_{S-1} \mid x_1, \ldots, x_{S-1})}{p(\lambda_S \mid x_1, \ldots, x_{S-1})}, \tag{137}$$

$$= \frac{p(\lambda_S \mid \lambda_{S-1})\, p(\lambda_{S-1} \mid x_1, \ldots, x_{S-1})}{\int d\lambda_{S-1}',\, p(\lambda_S \mid \lambda_{S-1}')\, p(\lambda_{S-1}' \mid x_1, \ldots, x_{S-1})}, \tag{138}$$

$$\propto \mathcal{N}\left(\lambda_S; \lambda_{S-1}, \sigma_\lambda^2\right) \mathcal{N}\left(\lambda_{S-1}; M_{S-1}^+, V_{S-1}^+\right), \tag{139}$$

$$= \mathcal{N}\left(\lambda_{S-1}; \frac{\lambda_S V_{S-1}^+ + M_{S-1}^+ \sigma_\lambda^2}{\sigma_\lambda^2 + V_{S-1}^+}, \frac{\sigma_\lambda^2 V_{S-1}^+}{\sigma_\lambda^2 + V_{S-1}^+}\right), \tag{140}$$

$$\propto \mathcal{N}\left(\lambda_S; \frac{\lambda_{S-1}(\sigma_\lambda^2 + V_{S-1}^+)}{V_{S-1}^+} - \frac{M_{S-1}^+ \sigma_\lambda^2}{V_{S-1}^+}, \frac{\sigma_\lambda^2(\sigma_\lambda^2 + V_{S-1}^+)}{V_{S-1}^+}\right). \tag{141}$$

We use standard results on normal distributions, enabling us to be a little shoddy with normalisations because the final result will be properly normalised. The final line above enables the convolution to be done easily:

$$p(\lambda_{S-1} \mid x_1, \ldots, x_S) \propto \mathcal{N}\left(M_S^-; \frac{\lambda_{S-1}(\sigma_\lambda^2 + V_{S-1}^+)}{V_{S-1}^+} - \frac{M_{S-1}^+ \sigma_\lambda^2}{V_{S-1}^+}, \frac{\sigma_\lambda^2(\sigma_\lambda^2 + V_{S-1}^+)}{V_{S-1}^+} + V_S^-\right), \tag{142}$$

$$= \mathcal{N}\left(\lambda_{S-1}; \frac{V_{S-1}^+ M_S^-}{\sigma_\lambda^2 + V_{S-1}^+} + \frac{M_{S-1}^+ \sigma_\lambda^2}{\sigma_\lambda^2 + V_{S-1}^+}, \frac{V_{S-1}^+}{\sigma_\lambda^2 + V_{S-1}^+}\left(\sigma_\lambda^2 + \frac{V_{S-1}^+ V_S^-}{\sigma_\lambda^2 + V_{S-1}^+}\right)\right), \tag{143}$$

$$= \mathcal{N}\left(\lambda_{S-1}; M_{S-1}^-, V_{S-1}^-\right). \tag{144}$$

K.4 The classical approach

Notice that the MAP estimate, $\hat{\lambda}_i$, of λ_i in the forward pass is just M_i^+. The classical approach is to write this as a "gain" like this:

$$\begin{aligned}
\hat{\lambda}_i &= M_i^+ && (145)\\
&= \frac{x_i P_i + M_{i-1}^+ \sigma_x^2}{P_i + \sigma_x^2} && (146)\\
&= \hat{\lambda}_{i-1} \frac{\sigma_x^2}{P_i + \sigma_x^2} + x_i \frac{P_i}{P_i + \sigma_x^2} && (147)\\
&= \hat{\lambda}_{i-1} \left(1 - \frac{P_i}{P_i + \sigma_x^2}\right) + x_i \frac{P_i}{P_i + \sigma_x^2} && (148)\\
&= \hat{\lambda}_{i-1} + \frac{P_i}{P_i + \sigma_x^2} \left(x_i - \hat{\lambda}_{i-1}\right). && (149)
\end{aligned}$$

The term

$$K_i = \frac{P_i}{P_i + \sigma_x^2} \tag{150}$$

is known as the Kalman gain. To my mind, the last line is daft; it obfuscates what is actually going on. The useful line is the one before, making it clear that the estimate is a linear combination of the previous state and new observation.

In the smoother, there is a similar term

$$J_i = \frac{V_{i-1}^+}{V_{i-1}^+ + \sigma_\lambda^2}, \tag{151}$$

that does a similar job to the Kalman gain.