The Kalman Filter

Max Welling

California Institute of Technology 136-93
Pasadena, CA 91125
welling@vision.caltech.edu

1 Introduction

Until now we have only looked at systems without any dynamics or structure over time. The models had spatial structure but were defined at one moment in time, i.e. they were "static". In this lecture we will analyse a powerfull dynamic model, which could be characterized as factor analysis over time, although the number of observed features is not necessarily larger than the number of factors. The idea is again to compute only the mean and the covariance statistics, i.e. to characterize the probabilities by Gaussians. This has the advantage of being completely tractible (for strongly nonlinear systems the Gaussian assumption can no longer hold). The power of the Kalman Filter (KF), is that it operates on-line. This implies, that to compute the best estimate of the state and its uncertainty, we can update the previous estimates by the new measurement. This implies that we don't have to consider all the previous data again, to compute the optimal estimates; we only need to consider the estimates from the previous time step and the new measurement.

So what are KFs usually used for, or what do they model? It is not hard to motivate the KF, because in practice it can be used for almost "everthing that moves". Popular applications include, navigation, guidance, radar tracking, sonar ranging, satellite orbit computation, stock prize prediction, etc. These applications can summarized as denoising, tracking and control.

It is used in all sorts of fields, like engineering, seismology, bioengineering, econometrics etc. For instance, when the Eagle landed on the moon, it did so with a KF. Also, gyroscopes in airplanes use KFs. And the list goes on and on.

The main idea is that we like to estimate a state of some sort (location and velocity of airplane) and its uncertainty. However, we do not directly observe these states. We only observe some measurements from an array of sensors, which are noisy. As an additional complication, the states evolve in time, also with its own noise or uncertainties. The question now becomes, how can we still optimally use our measurements to estimate the unobserved (hidden) states and their uncertainties.

In the following we will first describe the Kalman Filter and derive the KF equations. We assume that the parameters of the system are fixed (known). Then we derive the Kalman Smoother equations, which allow us to use measurements forward in time to help predict the state at the current time better. Because these estimates are usually less noisy than the if we used measurements up till the current time only, we say that we smooth the state estimates. Next, we will show how to employ the Kalman Filter and smoother equations to efficiently estimate the parameters of the model from training data. This will then be, not surprisingly, another instance of the EM algorithm. In the appendix you will find some usefull lemma's and matrix equalities, together with the derivation of the "lag-one smoother", which is needed for learning.

2 Introductory Example

Let me briefly describe an example. Consider a ship at sea which has lost its bearings. They need to estimate their current position using the positions of the stars. In the meantime the ships moves on on a wavy sea. The question becomes, how can we incorporate the measurement optimally, to estimate the ships location at sea. Let's assume that the model for the ship's location over time is given by,

$$\mathbf{y}_{t+1} = \mathbf{y}_t + \mathbf{c} + \mathbf{w}_t \tag{1}$$

This contains a drift term (constant velocity $\mathbf{c}$), and a noise term[1]. A noisy measurement is described by,

$$\mathbf{x}_t = \mathbf{y}_t + \mathbf{v}_t \tag{2}$$

Let's also assume we have some estimate $\hat{\mathbf{y}}_t$ of the location at some time t and some uncertainty σ_t^2. How does this change as the ships sails for one second. Of course, it will drift in the direction of its velocity by a distance $\mathbf{c}$. However, the uncertainty grows, due to the waves. This is expressed as,

$$\hat{\mathbf{y}}_{t+1} = \hat{\mathbf{y}}_t + \mathbf{c} \tag{3}$$

[1] The constant drift term will not be used in the main body of these classnotes

The uncertainty is given by,

$$\sigma_{t+1}^2 = \sigma_t^2 + \sigma_{\mathbf{w}}^2 \tag{4}$$

If, we do not do any measurements, the uncertainty in the position will keep growing, until we have no clue anymore as to where we are. If we add the information of a measurement, the final estimate is weighted average between the observed position and the previous estimate,

$$\hat{\mathbf{y}}'_{t+1} = \frac{\sigma_{\mathbf{v}}^2}{\sigma_{t+1}^2 + \sigma_{\mathbf{v}}^2}\hat{\mathbf{y}}_{t+1} + \frac{\sigma_{t+1}^2}{\sigma_{t+1}^2 + \sigma_{\mathbf{v}}^2}\mathbf{x}_{t+1} \tag{5}$$

We observe that if we have infinite confidence in the measurement $\sigma_{\mathbf{v}} \to 0$, then the new location estimate is simply equal to the measurement. Also, if we have infinite confidence in the previous estimate, the measurement is ignored. For the new uncertainty we find,

$$\sigma'^2_{t+1} = \frac{\sigma_{t+1}^2\ \sigma_{\mathbf{v}}^2}{\sigma_{t+1}^2 + \sigma_{\mathbf{v}}^2} \tag{6}$$

This is also easy to interpret, since it says that if one of the uncertainties dissappears, the total uncertainty disappears, since the other estimate can simply be ignored in that case. Notice that the uncertainty always decreases or stays equal, by adding the measurement. The estimates for the location and uncertainty, incorporatating the measurement, can be rewritten as follows,

$$\hat{\mathbf{y}}'_{t+1} = \hat{\mathbf{y}}_{t+1} + \mathbf{K}_{t+1}(\mathbf{x}_{t+1} - \hat{\mathbf{y}}_{t+1}) \tag{7}$$

$$\sigma'^2_{t+1} = (1 - \mathbf{K}_{t+1})\ \sigma_{t+1}^2 \tag{8}$$

$$\mathbf{K}_{t+1} = \frac{\sigma_{t+1}^2}{\sigma_{t+1}^2 + \sigma_{\mathbf{v}}^2} \tag{9}$$

From this we can see that the state estimate is corrected by the measurement error, multiplied by a gain factor (the Kalman gain). If the gain is zero, no attention is paid to the measurement, if its one, we simply use the measurement as our new state estimate. Similarly for the uncertainties, if the measurement is infinitely accurate, the gain is one, which implies that there is no uncertainty left. On the other hand, if the measurement is worthless, the gain is zero, which therefore does not decrease the overall uncertainty.

The above one dimensional example will be generalized to higher dimensions in the following.

3 The Model

Let us first introduce the state of the KF $\mathbf{y}_t$ at time t. The state is a vector of dimension d and remains unobserved. At every time t we alo have a k dimenional vector of observations $\mathbf{x}_t$, which depend on the state and some additive Gaussian noise. We will assume the following dynamical model for the KF:

$$\mathbf{y}_{t+1} = \mathbf{A}\mathbf{y}_t + \mathbf{w}_t \tag{10}$$

$$\mathbf{x}_t = \mathbf{B}\mathbf{y}_t + \mathbf{v}_t \tag{11}$$

Note that the dynamics is governed by a Markov process, i.e. the state at $\mathbf{y}_{t+1}$ is independent of all other states, given $\mathbf{y}_t$. The evolution noise and the measurement noise are assumed white and Gaussian, i.e. distributed according to,

$$\mathbf{w} \sim \mathcal{G}_{\mathbf{w}}[0, \mathbf{Q}] \tag{12}$$

$$\mathbf{v} \sim \mathcal{G}_{\mathbf{v}}[0, \mathbf{R}] \tag{13}$$

The noise vectors $\mathbf{v}_t$ and $\mathbf{w}_t$ are also assumed to be uncorrelated with the the state $\mathbf{y}_t$. From this we simply derive,

$$\mathbf{E}[\mathbf{y}_t, \mathbf{v}_k] = 0 \quad \forall \quad t, k \tag{14}$$

$$\mathbf{E}[\mathbf{y}_t, \mathbf{w}_k] = 0 \quad t \leq k \tag{15}$$

$$\mathbf{E}[\mathbf{x}_t, \mathbf{v}_k] = 0 \quad t \leq k - 1 \tag{16}$$

$$\mathbf{E}[\mathbf{x}_t, \mathbf{w}_k] = 0 \quad t \leq k \tag{17}$$

$$\mathbf{E}[\mathbf{v}_t, \mathbf{w}_k] = 0 \quad \forall \quad t, k \tag{18}$$

$$\mathbf{E}[\mathbf{v}_t, \mathbf{v}_k] = 0 \quad t \neq k, \qquad = R \quad t = k \tag{19}$$

$$\mathbf{E}[\mathbf{w}_t, \mathbf{w}_k] = 0 \quad t \neq k, \qquad = Q \quad t = k \tag{20}$$

The above model could be considered as a factor analysis model over time, i.e. at every instant we have a FA model, where the factors now depend on the factors of a previous time step.

The initial state $\mathbf{y}_1$ is distributed according to

$$\mathbf{y}_1 \sim \mathcal{G}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]. \tag{21}$$

It is easy to generalize this model to include a 'drift' and external inputs. The drift is a constant change expressed by adding $\boldsymbol{\nu}$ to the dynamical equation.

$$\begin{aligned} \mathbf{y}_{t+1} &= \mathbf{A}\mathbf{y}_t + \boldsymbol{\nu} + \mathbf{w}_t \\ &= \mathbf{A}'\mathbf{y}'_t + \mathbf{w}_t \end{aligned} \tag{22}$$

where we incorporated the constant again through the redefinitions $\mathbf{A}' = [\mathbf{A}, \boldsymbol{\nu}]$ and $\mathbf{y}'_t = [\mathbf{y}_t, 1]$. For the external inputs, we let the evolution depend on a l dimensional vector of inputs, $\mathbf{u}_t$, as follows,

$$\mathbf{y}_{t+1} = \mathbf{A}\mathbf{y}_t + \mathbf{C}\mathbf{u}_t + \mathbf{w}_t \tag{23}$$

This model is used when we want to control the system. We could even let the parameters $\{\mathbf{A}, \boldsymbol{\nu}, \mathbf{C}, \}$ depend on time. However, in the rest of this chapter we will assume the simplest case, i.e. a linear evolution without drift or inputs, and a linear measurement equation, with white uncorrelated noise. Since the initial state is Gaussian and the evolution equation is linear, this implies that the state at later times will remain Gaussian.

4 General Properties

We want to be able to estimate the state and the covariance of the state at any time t, given a set of observations $\mathbf{x}^\tau = \{\mathbf{x}_1, ..., \mathbf{x}_\tau\}$. If τ is equal to the current time t we say that we filter the state. If τ is smaller than t we say that we predict the state and finally, if τ is larger the t we say that we smooth the state. The main probability of interest is,

$$p(\mathbf{y}_t|\mathbf{x}^\tau), \tag{24}$$

since it conveys all the information about the state $\mathbf{y}_t$ at time t, given all the observations up to time τ. Since we are assuming that this probability is Gaussian, we only need to calculate its mean and covariance, denoted by

$$\hat{\mathbf{y}}_t^\tau = \mathbf{E}[\mathbf{y}_t|\mathbf{x}^\tau] \tag{25}$$

$$\mathbf{P}_t^\tau = \mathbf{E}[\tilde{\mathbf{y}}_t^\tau \tilde{\mathbf{y}}_t^\tau|\mathbf{x}^\tau] \tag{26}$$

where we defined $\tilde{\mathbf{y}}_t^\tau = \mathbf{y}_t - \hat{\mathbf{y}}_t^\tau$, i.e. the state prediction error. Notice that these quantities still depend on the random variables $\mathbf{x}^\tau$ and are therefore random variables themselves. We will now prove however that the covariance $\mathbf{P}$ does actually not depend on $\mathbf{x}^\tau$. $\mathbf{P}$ may be considered as a parameter therefore in the following. To proof the above claim we simply show that the correlation between the random variables $\tilde{\mathbf{y}}_t^\tau$ and $\mathbf{x}^\tau$ vanishes. For normally distributed random variables this implies that they are independent.

Lemma 3 The random variables $\tilde{\mathbf{y}}_t^\tau = \mathbf{y}_t - \hat{\mathbf{y}}_t^\tau$ and $\mathbf{x}^\tau = \{\mathbf{x}_1, ..., \mathbf{x}_\tau\}$ are independent.

proof

$$\begin{aligned} &\mathbf{E}[\mathbf{y}_t \mathbf{x}^\tau] - \mathbf{E}[\hat{\mathbf{y}}_t^\tau(\mathbf{x}^\tau)\ \mathbf{x}^\tau] = \\ &\int d\mathbf{y}_t d\mathbf{x}^\tau\ p(\mathbf{y}_t, \mathbf{x}^\tau)\ \mathbf{y}_t\ \mathbf{x}^\tau - \int d\mathbf{x}^\tau\ p(\mathbf{x}^\tau)\ \mathbf{x}^\tau \left[\int d\mathbf{y}_t\ p(\mathbf{y}_t|\mathbf{x}^\tau)\ \mathbf{y}_t\right] \\ &\int d\mathbf{y}_t d\mathbf{x}^\tau\ p(\mathbf{y}_t, \mathbf{x}^\tau)\ \mathbf{y}_t\ \mathbf{x}^\tau - \int d\mathbf{y}_t d\mathbf{x}^\tau\ p(\mathbf{x}^\tau)\ p(\mathbf{y}_t|\mathbf{x}^\tau)\ \mathbf{y}_t\ \mathbf{x}^\tau \\ &= 0 \quad \text{since} \quad p(\mathbf{y}_t, \mathbf{x}^\tau) = p(\mathbf{x}^\tau)\ p(\mathbf{y}_t|\mathbf{x}^\tau) \end{aligned}$$

From this we derive the following corollary,

$$\mathbf{P}_{t_1,t_2}^\tau = \mathbf{E}[\tilde{\mathbf{y}}_{t_1}^\tau \tilde{\mathbf{y}}_{t_2}^\tau] = \mathbf{E}[\mathbf{y}_{t_1}\mathbf{y}_{t_2}] - \mathbf{E}[\hat{\mathbf{y}}_{t_1}^\tau \hat{\mathbf{y}}_{t_2}^\tau] \tag{27}$$

Another usefull result we will need is the fact that the predicted measurement error $\boldsymbol{\varepsilon}_t = \mathbf{x}_t - \mathbf{B}\hat{\mathbf{y}}_t^{t-1}$ is independent of the measurements $\mathbf{x}^{t-1}$. The predicted measurement error is also called the innovation, since it represents that part of the new measurement $\mathbf{x}_t$ that can not be predicted using knowledge of the $\mathbf{x}^{t-1}$ measurements, since they are independent.

Lemma 4 The random variables $\varepsilon_t = \mathbf{x}_t - \mathbf{B}\hat{\mathbf{y}}_t^{t-1}$ and $\mathbf{x}^{t-1} = \{\mathbf{x}_1, ..., \mathbf{x}_{t-1}\}$ are independent.
Proof The proof is simple, and proceeds again by proving that they are uncorrelated,

$$\begin{aligned}
\mathbf{E}[\varepsilon_t, \mathbf{x}^{t-1}] &= \mathbf{E}[\mathbf{x}_t \mathbf{x}^{t-1}] - \mathbf{B}\mathbf{E}[\hat{\mathbf{y}}_t^{t-1} \mathbf{x}^{t-1}] \\
&= \mathbf{B}\mathbf{E}[\mathbf{y}_t \mathbf{x}^{t-1}] + \mathbf{B}\mathbf{E}[\mathbf{v}_t \mathbf{x}^{t-1}] - \mathbf{B}\mathbf{E}[\hat{\mathbf{y}}_t^{t-1} \mathbf{x}^{t-1}] \\
&= \mathbf{B}\mathbf{E}[\tilde{\mathbf{y}}_t^{t-1} \mathbf{x}^{t-1}] + \mathbf{B}\mathbf{E}[\mathbf{v}_t \mathbf{x}^{t-1}] \\
&= 0,
\end{aligned}$$

where we used the result that $\tilde{\mathbf{y}}_t^{t-1}$ is independent of $\mathbf{x}^{t-1}$ proved directly above and (16).
Notice, that since the innovation ε_{t-1} is a function of $\mathbf{x}^{t-1}$, this also implies the following corollary,

$$\mathbf{E}[\varepsilon_t \varepsilon_\tau] = 0 \qquad \text{for} \tau = 1, ..., t-1 \tag{28}$$

Before we proceed we would like to remark we have not well motivated $\hat{\mathbf{y}}_t^\tau$ as the preferred estimate of the state $\mathbf{y}_t$ given data $\mathbf{x}^\tau$. Other possible choices could be, the most likely state given the data, or the most likely sequence of states $\mathbf{y}_1, ..., \mathbf{y}_t$ given the data. It turns out that for Gaussian distributed random variables, these objectives are equivalent. On top of that, it turns out that the $\hat{\mathbf{y}}_t^\tau$ is also the minimal variance estimator, i.e. it minimizes $\mathbf{P}_t^\tau$.

5 Kalman Filter Equations

We will now proceed to derive the Kalman *filter* equations (i.e. $\tau = t$). First write,

$$p(\mathbf{y}_t|\mathbf{x}^t) = \frac{p(\mathbf{x}_t|\mathbf{y}_t)\ p(\mathbf{y}_t|\mathbf{x}^{t-1})}{p(\mathbf{x}_t|\mathbf{x}^{t-1})}, \tag{29}$$

where

$$p(\mathbf{y}_t|\mathbf{x}^{t-1}) = \int d\mathbf{y}_{t-1}\ p(\mathbf{y}_t|\mathbf{y}_{t-1})\ p(\mathbf{y}_{t-1}|\mathbf{x}^{t-1}) \tag{30}$$

The denominator in (29) is an unimportant normalization factor. The remaining densities are given by,

$$\begin{aligned}
p(\mathbf{x}_t|\mathbf{y}_t) &= \mathcal{G}_{\mathbf{x}_t}[\mathbf{B}\mathbf{y}_t, \mathbf{R}] &(31)\\
p(\mathbf{y}_t|\mathbf{y}_{t-1}) &= \mathcal{G}_{\mathbf{y}_t}[\mathbf{A}\mathbf{y}_{t-1}, \mathbf{Q}] &(32)\\
& &(33)
\end{aligned}$$

The equations (29) and (30) have interesting interpretations as a reactive reenforcement due to an observation and a diffusion equation for the Gaussian probability between observations. Equation (29) is basically applying Bayes law in the same way as we did for Bayesian learning, i.e. we are updating the probability distribution of an unknown random variable by including evidence. This has the effect of making the distribution peakier, i.e. less uncertain. The second equation (30) evolves the hidden state from one instant to the next, without considering more evidence. This has the effect of making the distribution less peakier, i.e. introducing more uncertainty. Together, these equations express $p(\mathbf{y}_t|\mathbf{x}^t)$ in terms of $p(\mathbf{y}_{t-1}|\mathbf{x}^{t-1})$ and may be used recursively. It is now easy to verify that in case of a prediction only the equation (30) remains, i.e.

$$p(\mathbf{y}_t|\mathbf{x}^\tau) = \int d\mathbf{y}_{t-1}\ p(\mathbf{y}_t|\mathbf{y}_{t-1})\ p(\mathbf{y}_{t-1}|\mathbf{x}^\tau) \quad \tau < t \tag{34}$$

The case when $\tau > t$, i.e. smoothing, is more difficult and will be dealt with later. Let us return to the filtering case. We like to calculate the mean $\hat{\mathbf{y}}_t^{t-1}$ and covariance $\mathbf{P}_t^{t-1}$ of the pdf $p(\mathbf{y}_t|\mathbf{x}^{t-1})$, expressed in terms of the mean $\hat{\mathbf{y}}_{t-1}^{t-1}$ and covariance $\mathbf{P}_{t-1}^{t-1}$ of the density $p(\mathbf{y}_{t-1}|\mathbf{x}^{t-1})$. Notice that these two estimators determine the densities completely, since they are Gaussian. For the mean we find,

$$\begin{aligned}
\hat{\mathbf{y}}_t^{t-1} &= \mathbf{E}[\mathbf{y}_t|\mathbf{x}^{t-1}] \\
&= \mathbf{A}\mathbf{E}[\mathbf{y}_{t-1}|\mathbf{x}^{t-1}] + \mathbf{E}[\mathbf{w}_{t-1}|\mathbf{x}^{t-1}] \\
&= \mathbf{A}\hat{\mathbf{y}}_{t-1}^{t-1} &(35)
\end{aligned}$$

where we have used (17) and the fact that $\mathbf{w}$ has zero mean (12). For the covariance we write we first write,

$$\begin{aligned}\tilde{\mathbf{y}}_t^{t-1} &= (\mathbf{y}_t - \hat{\mathbf{y}}_t^{t-1}) \\ &= \mathbf{A}\mathbf{y}_{t-1} + \mathbf{w}_{t-1} - \mathbf{A}\hat{\mathbf{y}}_{t-1}^{t-1} \\ &= \mathbf{A}\tilde{\mathbf{y}}_{t-1}^{t-1} + \mathbf{w}_{t-1}\end{aligned}$$

and notice that $\mathbf{w}_{t-1}$ is independent of $\tilde{\mathbf{y}}_{t-1}^{t-1}$ since this is a function of $\mathbf{y}_{t-1}$ and $\mathbf{x}^{t-1}$, while $\mathbf{w}_{t-1}$ is independent of both (using 15 and 17). Thus we can write,

$$\begin{aligned}\mathbf{P}_t^{t-1} &= \mathbf{E}[\tilde{\mathbf{y}}_t^{t-1}(\tilde{\mathbf{y}}_t^{t-1})^T] \\ &= \mathbf{E}[(\mathbf{A}\tilde{\mathbf{y}}_{t-1}^{t-1} + \mathbf{w}_{t-1})(\mathbf{A}\tilde{\mathbf{y}}_{t-1}^{t-1} + \mathbf{w}_{t-1})^T] \\ &= \mathbf{A}\mathbf{E}[\tilde{\mathbf{y}}_{t-1}^{t-1}(\tilde{\mathbf{y}}_{t-1}^{t-1})^T]\mathbf{A}^T + \mathbf{E}[\mathbf{w}_{t-1}\mathbf{w}_{t-1}^T] \\ &= \mathbf{A}\mathbf{P}_{t-1}^{t-1}\mathbf{A}^T + \mathbf{Q} \qquad (36)\end{aligned}$$

Next we wish to calculate $\hat{\mathbf{y}}_t^t$ and P_t^t interms of the above calculated quantities, using equation (29). We write,

$$p(\mathbf{y}_t|\mathbf{x}^t) = \frac{\mathcal{G}_{\mathbf{y}_t}[\hat{\mathbf{y}}_t^{t-1}, \mathbf{P}_t^{t-1}]\ \mathcal{G}_{\mathbf{x}_t}[\mathbf{B}\mathbf{y}_t, \mathbf{R}]}{p(\mathbf{x}_t|\mathbf{x}^{t-1})}. \qquad (37)$$

We will now use Lemma 1, applying (74) to the second term in the numerator of (37), and then applying (75) to the result of that we find,

$$p(\mathbf{y}_t|\mathbf{x}^t) = k_2(\mathbf{x}^t)\ \mathcal{G}_{\mathbf{y}_t}[([\mathbf{P}_t^{t-1}]^{-1} + \mathbf{B}^T\mathbf{R}^{-1}\mathbf{B})^{-1}([\mathbf{P}_t^{t-1}]^{-1}\hat{\mathbf{y}}_t^{t-1} + \mathbf{B}^T\mathbf{R}^{-1}\mathbf{x}_t), ([\mathbf{P}_t^{t-1}]^{-1} + \mathbf{B}^T\mathbf{R}^{-1}\mathbf{B})^{-1}], \qquad (38)$$

Because we know that the multplication of two Gaussians is again a Gaussian and moreover that (38) must be normalized with respect to $\mathbf{y}_t$, we deduce the the factor $k_2(\mathbf{x}^t) = 1$. Finally we must use Lemma 2 to show that $p(\mathbf{y}_t|\mathbf{x}^t)$ is a Gaussian distribution with the following mean and covariance,

$$\begin{aligned}\hat{\mathbf{y}}_t^t &= \hat{\mathbf{y}}_t^{t-1} + \mathbf{K}_t(\mathbf{x}_t - \mathbf{B}\hat{\mathbf{y}}_t^{t-1}) \qquad (39) \\ \mathbf{P}_t^t &= (\mathbf{I} - \mathbf{K}_t\mathbf{B})\mathbf{P}_t^{t-1} \qquad (40) \\ \mathbf{K}_t &= \mathbf{P}_t^{t-1}\mathbf{B}^T(\mathbf{R} + \mathbf{B}\mathbf{P}_t^{t-1}\mathbf{B}^T)^{-1} \qquad (41)\end{aligned}$$

where $\mathbf{K}_t$ is called the Kalman gain. These equations are initialized by $\hat{\mathbf{y}}_1^0 = \boldsymbol{\mu}$ and $\mathbf{P}_1^0 = \boldsymbol{\Sigma}$. These equations, together with (35) and (36) constitute the celebrated Kalman Filter equations and allow one to estimate the state of the system on-line, i.e. every new observation can be used recursively, given the information that was already received before. Notice that the gain factor $\mathbf{K}_t$ grows if the measurement covariance $\mathbf{R}$ becomes smaller, thus putting more weight on the measurement residual (difference between predicted an actual measurement). Also if the noise covariance $\mathbf{P}_t^{t-1}$ becomes smaller, less emphasis is put on the measurement residual. It is also instructive to notice that the evolution of the state noise $\mathbf{P}_t^t$ (and therefore the Kalman gain $\mathbf{K}_t$), is independent of the measurements and states and may be precomputed. From a numerical point of view, equation (40) is not preferable, due to the fact that it is a difference of two postive definite matrices, which is not guaranteed to result in a positive definite matrix, and may lead to numerical instabilities. This however, is easily fixed by noting that from (41) we can derive,

$$\mathbf{K}_t\mathbf{R}\mathbf{K}_t^T = (\mathbf{I} - \mathbf{K}_t\mathbf{B})\mathbf{P}_t^{t-1}\mathbf{B}^T\mathbf{K}^T, \qquad (42)$$

which can be used to rewrite (40) as,

$$\mathbf{P}_t^t = (\mathbf{I} - \mathbf{K}_t\mathbf{B})\mathbf{P}_t^{t-1}(\mathbf{I} - \mathbf{K}_t\mathbf{B})^T + \mathbf{K}_t\mathbf{R}\mathbf{K}_t^T, \qquad (43)$$

which is a sum of two postive definite matrices!

6 Kalman Smoother Equations

Next we want to solve the smoothing problem. This implies that we are going to include later measurements $\mathbf{x}_\tau, \quad \tau > t$, to improve our estimates of the states $\mathbf{y}_t$. The resultant estimates will be smoother (less noisy).

First concentrate on the mean $\mathbf{E}[\mathbf{y}_{t-1}|\mathbf{x}^\tau]$, for $\tau > t$. We will now invoke the corollary of Lemma 1, identifying $\mathbf{y} = \mathbf{y}_{t-1}$, $\mathbf{x} = \mathbf{y}_t$, $\boldsymbol{\mu}_y = \hat{\mathbf{y}}_{t-1}^\tau$, $\boldsymbol{\mu}_x = \hat{\mathbf{y}}_t^\tau$ and $\mathcal{G}_\mathbf{z}[\boldsymbol{\mu}, \boldsymbol{\Sigma}] = p(\mathbf{y}_{t-1}, \mathbf{y}_t|\mathbf{x}^\tau)$. Using these identifications we find,

$$\hat{\mathbf{y}}_{t-1}^\tau = \mathbf{E}[\mathbf{y}_{t-1}|\mathbf{x}^\tau] = \mathbf{E}[\mathbf{y}_{t-1}|\mathbf{y}_t = \hat{\mathbf{y}}_t^\tau, \mathbf{x}^\tau]. \tag{44}$$

Next we write,

$$\begin{aligned}
p(\mathbf{y}_{t-1}, \mathbf{y}_t|\mathbf{x}^\tau) &= \frac{p(\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{x}^{t-1}, \mathbf{x}_t, ..., \mathbf{x}_\tau)}{p(\mathbf{x}^\tau)} \\
&= \frac{p(\mathbf{x}_t, ..., \mathbf{x}_\tau|\mathbf{y}_{t-1}, \mathbf{y}_t, \mathbf{x}^{t-1})\ p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{x}^{t-1})\ p(\mathbf{y}_{t-1}|\mathbf{x}^{t-1})\ p(\mathbf{x}^{t-1})}{p(\mathbf{x}^\tau)} \\
&= \frac{p(\mathbf{x}_t, ..., \mathbf{x}_\tau|\mathbf{y}_t)\ p(\mathbf{y}_t|\mathbf{y}_{t-1})\ p(\mathbf{y}_{t-1}|\mathbf{x}^{t-1})}{p(\mathbf{x}_t, ..., \mathbf{x}_\tau|\mathbf{x}^{t-1})} \\
&= k_1(\mathbf{y}_t, \mathbf{x}^\tau)\ p(\mathbf{y}_t|\mathbf{y}_{t-1})\ p(\mathbf{y}_{t-1}|\mathbf{x}^{t-1}) \\
&= k_1(\mathbf{y}_t, \mathbf{x}^\tau)\ \mathcal{G}_{\mathbf{y}_t}[\mathbf{A}\mathbf{y}_{t-1}, \mathbf{Q}]\ \mathcal{G}_{\mathbf{y}_{t-1}}[\hat{\mathbf{y}}_{t-1}^{t-1}, \mathbf{P}_{t-1}^{t-1}]
\end{aligned} \tag{45}$$

In the same spirit as the derivation for the kalman filter (see derivation around 37), we will now use Lemma 1, applying (74) to the second term in (45), and then applying (75) to the result of that,

$$\begin{aligned}
&p(\mathbf{y}_{t-1}|\mathbf{y}_t, \mathbf{x}^\tau) = \frac{p(\mathbf{y}_{t-1}, \mathbf{y}_t|\mathbf{x}^\tau)}{p(\mathbf{y}_t|\mathbf{x}^\tau)} \\
&= k_2(\mathbf{y}_t, \mathbf{x}^\tau)\ \mathcal{G}_{\mathbf{y}_{t-1}}[([\mathbf{P}_{t-1}^{t-1}]^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A})^{-1}([\mathbf{P}_{t-1}^{t-1}]^{-1}\hat{\mathbf{y}}_{t-1}^{t-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{y}_t), ([\mathbf{P}_{t-1}^{t-1}]^{-1} + \mathbf{A}^T\mathbf{Q}^{-1}\mathbf{A})^{-1}]
\end{aligned} \tag{46}$$

Notice the similarity between (38) and (46). Because we know that the multplication of two Gaussians is again a Gaussian and moreover that (46) must be normalized with respect to $\mathbf{y}_{t-1}$, we deduce the the factor $k_4(\mathbf{y}_t, \mathbf{x}^\tau) = 1$. Finally we invoke (44) and Lemma 2 to find,

$$\hat{\mathbf{y}}_{t-1}^\tau = \hat{\mathbf{y}}_{t-1}^{t-1} + \mathbf{J}_{t-1}(\hat{\mathbf{y}}_t^\tau - \hat{\mathbf{y}}_t^{t-1}) \tag{47}$$

$$\mathbf{J}_{t-1} = \mathbf{P}_{t-1}^{t-1}\mathbf{A}^T[\mathbf{P}_t^{t-1}]^{-1} \tag{48}$$

where we used (36) in the last line. This is initialized with $\hat{\mathbf{y}}_\tau^\tau$, computed from the Kalman Filter equations. For the covariance we first observe,

$$\tilde{\mathbf{y}}_{t-1}^\tau + \mathbf{J}_{t-1}\hat{\mathbf{y}}_t^\tau = \tilde{\mathbf{y}}_{t-1}^{t-1} + \mathbf{J}_{t-1}\mathbf{A}\hat{\mathbf{y}}_{t-1}^{t-1}, \tag{49}$$

where we used (35) and (47). Multiplying both sides with their respective transpose from the right, taking expectations, and using the fact that $\tilde{\mathbf{y}}_{t-1}^\tau$ is independent of $\hat{\mathbf{y}}_t^\tau$ (since the latter is a function of $\mathbf{x}^\tau$ and Lemma 3) and similarly for $\tilde{\mathbf{y}}_{t-1}^{t-1}$ and $\hat{\mathbf{y}}_t^{t-1}$, gives us,

$$\mathbf{P}_{t-1}^\tau + \mathbf{J}_{t-1}\mathbf{E}[\hat{\mathbf{y}}_t^\tau\hat{\mathbf{y}}_t^\tau]\mathbf{J}_{t-1}^T = \mathbf{P}_{t-1}^{t-1} + \mathbf{J}_{t-1}\mathbf{A}\mathbf{E}[\hat{\mathbf{y}}_{t-1}^{t-1}\hat{\mathbf{y}}_{t-1}^{t-1}]\mathbf{A}^T\mathbf{J}_{t-1}^T. \tag{50}$$

Then we use,

$$\begin{aligned}
\mathbf{E}[\hat{\mathbf{y}}_t^\tau\hat{\mathbf{y}}_t^\tau] &= \mathbf{E}[\mathbf{y}_t\mathbf{y}_t] - \mathbf{P}_t^\tau &(51)\\
&= \mathbf{E}[(\mathbf{A}\mathbf{y}_{t-1} + \mathbf{w}_{t-1})(\mathbf{A}\mathbf{y}_{t-1} + \mathbf{w}_{t-1})^T] - \mathbf{P}_t^\tau &(52)\\
&= \mathbf{A}\mathbf{E}[\mathbf{y}_{t-1}\mathbf{y}_{t-1}]\mathbf{A}^T + \mathbf{Q} - \mathbf{P}_t^\tau &(53)
\end{aligned}$$

where (15) and (20) and the corollary following Lemma 3 was used. Analoguesly,

$$\mathbf{E}[\hat{\mathbf{y}}_{t-1}^{t-1}\hat{\mathbf{y}}_{t-1}^{t-1}] = \mathbf{E}[\mathbf{y}_{t-1}\mathbf{y}_{t-1}] - \mathbf{P}_{t-1}^{t-1} \tag{54}$$

Putting these together and using (36), we find

$$\mathbf{P}_{t-1}^\tau = \mathbf{P}_{t-1}^{t-1} + \mathbf{J}_{t-1}(\mathbf{P}_t^\tau - \mathbf{P}_t^{t-1})\mathbf{J}_{t-1}^T \tag{55}$$

which is initialized by $\mathbf{P}_\tau^\tau$, computed from the Kalman Filter equations. Equations (47), (55) and (48) are the so called Kalman smoother equations. If we want to estimate a state $\mathbf{y}_t$ given data $\mathbf{x}^\tau$ with $\tau > t$, then we first apply the Kalman filter equations recursively until we have reached the state $\mathbf{y}_\tau$. While moving forward we store the values for $\hat{\mathbf{y}}_t^t$, $\hat{\mathbf{y}}_t^{t-1}$, $\mathbf{P}_t^t$ and $\mathbf{P}_t^{t-1}$, $t = 1...\tau$. Then we move backward by applying the smoother equations, until we have reached the state t we would like to estimate. Because we include more observations in the estimation of the state, the result will be less noisy as compared to the Kalman filter result, hence the name smoother.

7 Parameter Estimation for the Kalman Model

We will now proceed to estimate the parameters $\{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{B}, \mathbf{R}, \mathbf{A}, \mathbf{Q}\}$ of the Kalman filter model using EM. We consider the states $\mathbf{y}_t$ as hidden variables, while $\mathbf{x}^\tau$ are the observations. We assume we have observed N sequences of length τ. The joint probability of the complete data is given by,

$$p(\mathbf{y}^\tau, \mathbf{x}^\tau) = p(\mathbf{y}_1) \prod_{t=2}^{\tau} p(\mathbf{y}_t|\mathbf{y}_{t-1}) \prod_{t=1}^{\tau} p(\mathbf{x}_t|\mathbf{y}_t) \tag{56}$$

For EM, we are interested in the expectation of the joint pdf over the posterior density,

$$\begin{aligned} Q &= \sum_{n=1}^{N} \int d\mathbf{y}^\tau \; p(\mathbf{y}^\tau|\mathbf{x}_n^\tau) \log\left[p(\mathbf{y}^\tau, \mathbf{x}_n^\tau)\right] \\ &= -\frac{1}{2} \sum_{n=1}^{N} \int d\mathbf{y}^\tau \; p(\mathbf{y}^\tau|\mathbf{x}_n^\tau)[(d+k)\tau \log(2\pi) \\ &+ \log\det\boldsymbol{\Sigma} + (\tau-1)\log\det\mathbf{Q} + \tau\log\det\mathbf{R} \\ &+ (\mathbf{y}_1 - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y}_1 - \boldsymbol{\mu}) \\ &+ \sum_{t=2}^{\tau} (\mathbf{y}_t - \mathbf{A}\mathbf{y}_{t-1})^T \mathbf{Q}^{-1} (\mathbf{y}_t - \mathbf{A}\mathbf{y}_{t-1}) \\ &+ \sum_{t=1}^{\tau} (\mathbf{x}_{t,n} - \mathbf{B}\mathbf{y}_t)^T \mathbf{R}^{-1} (\mathbf{x}_{t,n} - \mathbf{B}\mathbf{y}_t)] \end{aligned} \tag{57}$$

Inspection of this objective function reveals that the only sufficient statistics that need to be calculated in the E-step are,

$$\mathbf{E}[\mathbf{y}_t|\mathbf{x}_n^\tau] = \hat{\mathbf{y}}_{t,n}^\tau \qquad t = 1, ..., \tau \tag{58}$$

$$\mathbf{E}[\mathbf{y}_t\mathbf{y}_t|\mathbf{x}_n^\tau] = \mathbf{P}_t^\tau + \hat{\mathbf{y}}_{t,n}^\tau \hat{\mathbf{y}}_{t,n}^\tau \equiv \mathbf{M}_t^n \qquad t = 1, ..., \tau \tag{59}$$

$$\mathbf{E}[\mathbf{y}_t\mathbf{y}_{t-1}|\mathbf{x}_n^\tau] = \mathbf{P}_{t,t-1}^\tau + \hat{\mathbf{y}}_{t,n}^\tau \hat{\mathbf{y}}_{t-1,n}^\tau \equiv \mathbf{M}_{t,t-1}^n \qquad t = 2, ..., \tau \tag{60}$$

Fortunately, except for the last one, these are precisely the quantities that can be calculated through the Kalman Filter and Smoother recursions. The last quantity, which is called the *lag-one covariance smoother* is computed in the appendix C. It is given by the following recursion,

$$\mathbf{P}_{t-1,t-2}^\tau = \mathbf{P}_{t-1}^{t-1}\mathbf{J}_{t-2}^T + \mathbf{J}_{t-1}(\mathbf{P}_{t,t-1}^\tau - \mathbf{A}\mathbf{P}_{t-1}^{t-1})\mathbf{J}_{t-2}^T. \tag{61}$$

which is initialized by,

$$\mathbf{P}_{\tau,\tau-1}^\tau = (\mathbf{I} - \mathbf{K}_\tau\mathbf{B})\mathbf{A}\mathbf{P}_{\tau-1}^{\tau-1}. \tag{62}$$

We now concentrate on the M-step, which maximizes (57) with respect to the parameters of the joint density only, i.e. the parameters present in the posterior are held fixed.

Taking derivatives with respect to $\boldsymbol{\mu}$ and equating to zero gives,

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\mu}} Q = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\hat{\mathbf{y}}_{1,n}^\tau - \boldsymbol{\mu}) = 0 \Rightarrow \\ \boldsymbol{\mu}_{\mathbf{new}} = \frac{1}{N}\sum_{n=1}^{N} \hat{\mathbf{y}}_{1,n}^\tau \end{aligned} \tag{63}$$

For $\boldsymbol{\Sigma}$ this implies,

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\Sigma}^{-1}} Q = \frac{1}{2}N\boldsymbol{\Sigma} - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{M}_1^n - \hat{\mathbf{y}}_{1,n}^\tau\boldsymbol{\mu}^T - \boldsymbol{\mu}(\hat{\mathbf{y}}_{1,n}^\tau)^T + \boldsymbol{\mu}\boldsymbol{\mu}^T) \Rightarrow \\ \boldsymbol{\Sigma}_{\mathbf{new}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{M}_1^n - \boldsymbol{\mu}_{\mathbf{new}}\boldsymbol{\mu}_{\mathbf{new}}^T = \mathbf{P}_1^\tau + \frac{1}{N}\sum_{n=1}^{N}(\hat{\mathbf{y}}_{1,n}^\tau - \boldsymbol{\mu}_{\mathbf{new}})(\hat{\mathbf{y}}_{1,n}^\tau - \boldsymbol{\mu}_{\mathbf{new}})^T \end{aligned} \tag{64}$$

where we used that $\mathbf{P}_t^\tau$ is independent of $\mathbf{x}_n^\tau$. Taking derivatives with respect to $\mathbf{A}$ gives.

$$\frac{\partial}{\partial \mathbf{A}} Q = \sum_{n=1}^{N}\sum_{t=2}^{\tau}(\mathbf{Q}^{-1}\mathbf{M}_{t,t-1}^n - \mathbf{Q}^{-1}\mathbf{A}\mathbf{M}_{t-1}^n) \Rightarrow$$

$$\mathbf{A_{new}} = \left[\sum_{n=1}^{N}\sum_{t=2}^{\tau}\mathbf{M}_{t,t-1}^n\right]\left[\sum_{n=1}^{N}\sum_{t=2}^{\tau}\mathbf{M}_{t-1}^n\right]^{-1} \tag{65}$$

For $\mathbf{Q}$ we find,

$$\frac{\partial}{\partial \mathbf{Q}} Q = \frac{1}{2}(\tau-1)N\mathbf{Q} - \frac{1}{2}\sum_{n=1}^{N}\sum_{t=2}^{\tau}(\mathbf{M}_t^n - \mathbf{A}\mathbf{M}_{t-1,t}^n - \mathbf{M}_{t,t-1}\mathbf{A}^T + \mathbf{A}\mathbf{M}_{t-1}^n\mathbf{A}^T) \Rightarrow$$

$$\mathbf{Q_{new}} = \frac{1}{N(\tau-1)}\sum_{n=1}^{N}\sum_{t=2}^{\tau}(\mathbf{M}_t^n - \mathbf{A_{new}}\mathbf{M}_{t-1,t}^n), \tag{66}$$

where we used (65), and $\mathbf{M}_{t-1,t}^n = (\mathbf{M}_{t,t-1}^n)^T$. For $\mathbf{B}$ we have,

$$\frac{\partial}{\partial \mathbf{B}} Q = \sum_{n=1}^{N}\sum_{t=1}^{\tau}(\mathbf{R}^{-1}\mathbf{x}_{t,n}\hat{\mathbf{y}}_{t,n}^\tau - \mathbf{R}^{-1}\mathbf{B}\mathbf{M}_t^n) \Rightarrow$$

$$\mathbf{B_{new}} = \left[\sum_{n=1}^{N}\sum_{t=1}^{\tau}\mathbf{x}_{t,n}\hat{\mathbf{y}}_{t,n}^\tau\right]\left[\sum_{n=1}^{N}\sum_{t=1}^{\tau}\mathbf{M}_t^n\right]^{-1} \tag{67}$$

And finally we have for $\mathbf{R}$,

$$\frac{\partial}{\partial \mathbf{R}} Q = \frac{1}{2}\tau N\mathbf{Q} - \frac{1}{2}\sum_{n=1}^{N}\sum_{t=1}^{\tau}(\mathbf{x}_{t,n}\mathbf{x}_{t,n} - \mathbf{x}_{t,n}\hat{\mathbf{y}}_{t,n}^\tau\mathbf{B}^T - \mathbf{B}\hat{\mathbf{y}}_{t,n}^\tau\mathbf{x}_{t,n} + \mathbf{B}\mathbf{M}_t^n\mathbf{B}^T) \Rightarrow$$

$$\mathbf{R_{new}} = \frac{1}{N\tau}\sum_{n=1}^{N}\sum_{t=1}^{\tau}(\mathbf{x}_{t,n}\mathbf{x}_{t,n} - \mathbf{B_{new}}\hat{\mathbf{y}}_{t,n}^\tau\mathbf{x}_{t,n}), \tag{68}$$

where we used (67). Alternating E-steps and M-steps will thus converge to the maximum likelihood estimates of these parameters. Notice that the recursion equations fulfil a double role. They may be used to efficiently compute the E-step in the learning problem, and, once the parameters are fixed, to estimate the optimal state and state-covariance of the dynamical system, possibly on line.

8 Computation of Likelihood

To monitor the total log-likelihood of the system we may calculate,

$$\mathbf{L}_\tau = \sum_{n=1}^{N}\log[p(\mathbf{x}_n^\tau)] = \sum_{n=1}^{N}\sum_{t=2}^{\tau}\log[p(\mathbf{x}_{t,n}|\mathbf{x}_n^{t-1})] + \sum_{n=1}^{N}\log[p(\mathbf{x}_{1,n})]. \tag{69}$$

The mean and covariance of the Gaussians $p(\mathbf{x}_{t,n}|\mathbf{x}_{t-1,n})$ can be computed as follows (omitting n for notational convenience),

$$\begin{aligned}
\hat{\mathbf{x}}_t^{t-1} &= \int d\mathbf{x}_t\ p(\mathbf{x}_t|\mathbf{x}^{t-1})\ \mathbf{x}_t \\
&= \int d\mathbf{x}_t \int d\mathbf{y}_t\ p(\mathbf{x}_t|\mathbf{y}_t)\ p(\mathbf{y}_t|\mathbf{x}^{t-1})\ \mathbf{x}_t \\
&= \int d\mathbf{x}_t \int d\mathbf{y}_t\ \mathcal{G}_{\mathbf{x}_t}[\mathbf{B}\mathbf{y}_t, \mathbf{R}]\ \mathcal{G}_{\mathbf{y}_t}[\hat{\mathbf{y}}_t^{t-1}, \mathbf{P}_t^{t-1}]\ \mathbf{x}_t \\
&= \int d\mathbf{y}_t\ \mathcal{G}_{\mathbf{y}_t}[\hat{\mathbf{y}}_t^{t-1}, \mathbf{P}_t^{t-1}]\ \mathbf{B}\mathbf{y}_t \\
&= \mathbf{B}\hat{\mathbf{y}}_t^{t-1}.
\end{aligned} \tag{70}$$

For $p(\mathbf{x}_1)$ the above calculation gives,

$$\hat{\mathbf{x}}_1^0 = \mathbf{B}\boldsymbol{\mu} \tag{71}$$

Similarly, for the covariance we find,

$$\begin{aligned}
\mathbf{H}_t^{t-1} &= \int d\mathbf{x}_t \; p(\mathbf{x}_t|\mathbf{x}^{t-1}) \; (\mathbf{x}_t\mathbf{x}_t - \hat{\mathbf{x}}_t^{t-1}\hat{\mathbf{x}}_t^{t-1}) \\
&= \int d\mathbf{x}_t \int d\mathbf{y}_t \; \mathcal{G}_{\mathbf{x}_t}[\mathbf{B}\mathbf{y}_t, \mathbf{R}] \; \mathcal{G}_{\mathbf{y}_t}[\hat{\mathbf{y}}_t^{t-1}, \mathbf{P}_t^{t-1}] \; (\mathbf{x}_t\mathbf{x}_t - \hat{\mathbf{x}}_t^{t-1}\hat{\mathbf{x}}_t^{t-1}) \\
&= \int d\mathbf{y}_t \; \mathcal{G}_{\mathbf{y}_t}[\hat{\mathbf{y}}_t^{t-1}, \mathbf{P}_t^{t-1}] \; (\mathbf{R} + \mathbf{B}\mathbf{y}_t\mathbf{y}_t\mathbf{B}^T - \hat{\mathbf{x}}_t^{t-1}\hat{\mathbf{x}}_t^{t-1}) \\
&= \mathbf{R} + \mathbf{B}(\mathbf{P}_t^{t-1} + \hat{\mathbf{y}}_t^{t-1}\hat{\mathbf{y}}_t^{t-1})\mathbf{B}^T - \mathbf{B}\hat{\mathbf{y}}_t^{t-1}\hat{\mathbf{y}}_t^{t-1}\mathbf{B}^T \\
&= \mathbf{R} + \mathbf{B}\mathbf{P}_t^{t-1}\mathbf{B}^T.
\end{aligned} \tag{72}$$

The covariance for $p(\mathbf{x}_1)$ is then given by,

$$\mathbf{H}_1^0 = \mathbf{R} + \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T \tag{73}$$

A Lemma's

Lemma 1 Let $\mathcal{G}_{\mathbf{y}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]$ denote a normal density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. We have the following identities,

$$\mathcal{G}_{\mathbf{x}}[\mathbf{A}\mathbf{y}, \boldsymbol{\Sigma}] = k_1(\mathbf{x}) \; \mathcal{G}_{\mathbf{y}}[(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}, (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}], \tag{74}$$

$$\mathcal{G}_{\mathbf{y}}[\mathbf{a}, \mathbf{A}] \; \mathcal{G}_{\mathbf{y}}[\mathbf{b}, \mathbf{B}] = \mathcal{G}_{\mathbf{y}}[(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b}), (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}] \; \mathcal{G}_{\mathbf{a}}[\mathbf{b}, \mathbf{A} + \mathbf{B}]. \tag{75}$$

Also, if we write $\mathbf{z} = [\mathbf{y}, \mathbf{x}]$, $\boldsymbol{\mu} = [\boldsymbol{\mu}_y, \boldsymbol{\mu}_x]$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}$, then we have,

$$\mathbf{y} \sim \int d\mathbf{x} \; \mathcal{G}_{\mathbf{z}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}] = \mathcal{G}_{\mathbf{y}}[\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_{yy}] \tag{76}$$

$$\mathbf{y}|\mathbf{x} \sim \frac{\mathcal{G}_{\mathbf{z}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}]}{\mathcal{G}_{\mathbf{x}}[\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}]} = \mathcal{G}_{\mathbf{y}}[\boldsymbol{\mu}_y - \boldsymbol{\Sigma}_{yx}(\boldsymbol{\Sigma}_{xx})^{-1}(\boldsymbol{\mu}_x - \mathbf{x}), \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}(\boldsymbol{\Sigma}_{xx})^{-1}\boldsymbol{\Sigma}_{xy}] \tag{77}$$

As a corollary we notice that

$$\mathbf{E}[\mathbf{y}|\mathbf{x} = \boldsymbol{\mu}_x] = \mathbf{E}[\mathbf{y}] \tag{78}$$

Lemma 2 Consider a $d \times d$ matrix $\mathbf{P} > 0$, a $k \times k$ matrix $\mathbf{R} > 0$ and a $k \times d$ matrix $\mathbf{B}$, where $P > 0$ implies $\mathbf{a}^T\mathbf{P}\mathbf{a} > 0 \forall \;\; \mathbf{a}$ (i.e. positive eigenvalues). The following equalities hold,

$$(\mathbf{P}^{-1} + \mathbf{B}^T\mathbf{R}^{-1}\mathbf{B})^{-1} = \mathbf{P} - \mathbf{P}\mathbf{B}^T(\mathbf{B}\mathbf{P}\mathbf{B}^T + \mathbf{R})^{-1}\mathbf{B}\mathbf{P} \tag{79}$$

$$(\mathbf{P}^{-1} + \mathbf{B}^T\mathbf{R}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{R}^{-1} = \mathbf{P}\mathbf{B}^T(\mathbf{B}\mathbf{P}\mathbf{B}^T + \mathbf{R})^{-1} \tag{80}$$

B Matrix Identities

In the derivations to follow, the following idetities are useful,

$$\mathbf{a}^T\mathbf{A}\mathbf{b} = \mathbf{tr}[\mathbf{A}\mathbf{b}\mathbf{a}^T] \tag{81}$$

$$\mathbf{tr}[\mathbf{A}\mathbf{B}] = \mathbf{tr}[\mathbf{B}\mathbf{A}] \tag{82}$$

$$\frac{\partial}{\partial \mathbf{A}}\mathbf{tr}[\mathbf{A}\mathbf{B}] = \mathbf{B}^T \tag{83}$$

$$\frac{\partial}{\partial \mathbf{A}}\mathbf{tr}[\mathbf{A}^T\mathbf{B}] = \mathbf{B} \tag{84}$$

$$\log\det[\mathbf{A}] = -\log\det[\mathbf{A}^{-1}] \tag{85}$$

$$\frac{\partial}{\partial \mathbf{A}}\log\det[\mathbf{A}] = (\mathbf{A}^T)^{-1} \tag{86}$$

C Lag-One Covariance Smoother

In the E-step of the EM algorithm that estimated the parameters of the KF, we needed the filter and smoother equations. However, one quantity remains undetermined, which is the so called *lag-one-covariance-smoother*, $\mathbf{P}_{t,t-1}^{\tau}$. This quantity is only needed backwards. First we will derive the initial value the backward recursions,

$$\begin{aligned}
\mathbf{P}_{t,t-1}^{t} &= \mathbf{E}[\tilde{\mathbf{y}}_t^t \tilde{\mathbf{y}}_{t-1}^t] && (87)\\
&= \mathbf{E}[\{\tilde{\mathbf{y}}_t^{t-1} - \mathbf{K}_t(\mathbf{x}_t - \mathbf{B}\hat{\mathbf{y}}_t^{t-1})\}\{\tilde{\mathbf{y}}_{t-1}^{t-1} - \mathbf{J}_{t-1}\mathbf{K}_t(\mathbf{x}_t - \mathbf{B}\hat{\mathbf{y}}_t^{t-1})\}] && (88)\\
&= \mathbf{E}[\{\tilde{\mathbf{y}}_t^{t-1} - \mathbf{K}_t(\mathbf{B}\tilde{\mathbf{y}}_t^{t-1} + \mathbf{v}_t)\}\{\tilde{\mathbf{y}}_{t-1}^{t-1} - \mathbf{J}_{t-1}\mathbf{K}_t(\mathbf{B}\tilde{\mathbf{y}}_t^{t-1} + \mathbf{v}_t)\}] && (89)
\end{aligned}$$

where we used (11), (39) and (47) with $\tau = t$. Next, we write this out and use that both $\tilde{\mathbf{y}}_{t-1}^{t-1}$ and $\tilde{\mathbf{y}}_t^{t-1}$ are independent of $\mathbf{v}_t$, which is proved using (14) and (16). This will give

$$\begin{aligned}
\mathbf{P}_{t,t-1}^{t} &= \mathbf{P}_{t,t-1}^{t-1} - \mathbf{P}_t^{t-1}\mathbf{B}^T\mathbf{K}_t^T\mathbf{J}_{t-1}^T - \mathbf{K}_t\mathbf{B}\mathbf{P}_{t,t-1}^{t-1} + \mathbf{K}_t(\mathbf{B}\mathbf{P}_t^{t-1}\mathbf{B}^T + \mathbf{R})\mathbf{K}_t^T\mathbf{J}_{t-1}^T && (90)\\
&= (\mathbf{I} - \mathbf{K}_t\mathbf{B})\mathbf{A}\mathbf{P}_{t-1}^{t-1}, && (91)
\end{aligned}$$

where in the last line we used (41) and $\mathbf{P}_{t,t-1}^{t-1} = \mathbf{A}\mathbf{P}_{t-1}^{t-1}$, which is proved using the fact that $\mathbf{w}_{t-1}$ is independent of $\tilde{\mathbf{y}}_{t-1}^{t-1}$ (15, 17). If we set $t = \tau$ we find the initial condition,

$$\mathbf{P}_{\tau,\tau-1}^{\tau} = (\mathbf{I} - \mathbf{K}_\tau\mathbf{B})\mathbf{A}\mathbf{P}_{\tau-1}^{\tau-1}. \quad (92)$$

The derivation for the backward recursion is somewhat elaborate. First we write from (47) and (35),

$$\begin{aligned}
\tilde{\mathbf{y}}_{t-1}^{\tau} + \mathbf{J}_{t-1}\hat{\mathbf{y}}_t^{\tau} &= \tilde{\mathbf{y}}_{t-1}^{t-1} + \mathbf{J}_{t-1}\mathbf{A}\hat{\mathbf{y}}_{t-1}^{t-1} && (93)\\
\tilde{\mathbf{y}}_{t-2}^{\tau} + \mathbf{J}_{t-2}\hat{\mathbf{y}}_{t-1}^{\tau} &= \tilde{\mathbf{y}}_{t-2}^{t-2} + \mathbf{J}_{t-2}\mathbf{A}\hat{\mathbf{y}}_{t-2}^{t-2} && (94)
\end{aligned}$$

Next, we equate,

$$\begin{aligned}
&\mathbf{E}[(\hat{\mathbf{y}}_{t-1}^{\tau} + \mathbf{J}_{t-1}\hat{\mathbf{y}}_t^{\tau})(\hat{\mathbf{y}}_{t-2}^{\tau} + \mathbf{J}_{t-2}\hat{\mathbf{y}}_{t-2}^{\tau})^T] =\\
&\mathbf{E}[(\hat{\mathbf{y}}_{t-1}^{t-1} + \mathbf{J}_{t-1}\mathbf{A}\hat{\mathbf{y}}_{t-1}^{t-1})(\hat{\mathbf{y}}_{t-2}^{t-2} + \mathbf{J}_{t-2}\mathbf{A}\hat{\mathbf{y}}_{t-2}^{t-2})^T] \Rightarrow && (95)\\
&\mathbf{P}_{t-1,t-2}^{\tau} + \mathbf{J}_{t-1}\mathbf{E}[\hat{\mathbf{y}}_t^{\tau}\hat{\mathbf{y}}_{t-1}^{\tau}]\mathbf{J}_{t-2}^T =\\
&\mathbf{E}[\tilde{\mathbf{y}}_{t-1}^{t-1}\tilde{\mathbf{y}}_{t-2}^{t-2}] + \mathbf{J}_{t-1}\mathbf{A}\mathbf{E}[\hat{\mathbf{y}}_{t-1}^{t-1}\tilde{\mathbf{y}}_{t-2}^{t-2}] + \mathbf{J}_{t-1}\mathbf{A}\mathbf{E}[\hat{\mathbf{y}}_{t-1}^{t-1}\hat{\mathbf{y}}_{t-2}^{t-2}]\mathbf{A}^T\mathbf{J}_{t-2}^T, && (96)
\end{aligned}$$

where lemma 3 was used several times to get rid of cross-terms. We will now rewrite some of the terms appearing above,

$$\begin{aligned}
&\mathbf{E}[\hat{\mathbf{y}}_t^{\tau}\hat{\mathbf{y}}_{t-1}^{\tau}] = \mathbf{E}[\mathbf{y}_t\mathbf{y}_{t-1}] - \mathbf{P}_{t,t-1}^{\tau} =\\
&\mathbf{E}[(\mathbf{A}\mathbf{y}_{t-1} + \mathbf{w}_{t-1})(\mathbf{A}\mathbf{y}_{t-2} + \mathbf{w}_{t-2})^T] - \mathbf{P}_{t,t-1}^{\tau} =\\
&\mathbf{A}\mathbf{E}[\mathbf{y}_{t-1}\mathbf{y}_{t-2}]\mathbf{A}^T + \mathbf{A}\mathbf{E}[\mathbf{y}_{t-1}\mathbf{w}_{t-2}] - \mathbf{P}_{t,t-1}^{\tau} =\\
&\mathbf{A}\mathbf{E}[\mathbf{y}_{t-1}\mathbf{y}_{t-2}]\mathbf{A}^T + \mathbf{A}\mathbf{E}[(\mathbf{A}\mathbf{y}_{t-2} + \mathbf{w}_{t-2})\mathbf{w}_{t-2}] - \mathbf{P}_{t,t-1}^{\tau} =\\
&\mathbf{A}\mathbf{E}[\mathbf{y}_{t-1}\mathbf{y}_{t-2}]\mathbf{A}^T + \mathbf{A}\mathbf{Q} - \mathbf{P}_{t,t-1}^{\tau}. && (97)
\end{aligned}$$

Next,

$$\begin{aligned}
&\mathbf{E}[\tilde{\mathbf{y}}_{t-1}^{t-1}\tilde{\mathbf{y}}_{t-2}^{t-2}] = \mathbf{E}[\{\tilde{\mathbf{y}}_{t-1}^{t-2} - \mathbf{K}_{t-1}(\mathbf{x}_{t-1} - \mathbf{B}\hat{\mathbf{y}}_{t-1}^{t-2})\}\tilde{\mathbf{y}}_{t-2}^{t-2}] =\\
&\mathbf{P}_{t-1,t-2}^{t-2} - \mathbf{K}_{t-1}\mathbf{E}[(\mathbf{B}\tilde{\mathbf{y}}_{t-1}^{t-2} + \mathbf{v}_{t-1})\tilde{\mathbf{y}}_{t-2}^{t-2}] =\\
&\mathbf{P}_{t-1,t-2}^{t-2} - \mathbf{K}_{t-1}\mathbf{B}\mathbf{P}_{t-1,t-2}^{t-2}, && (98)
\end{aligned}$$

where (39) and (11) was used. Next,

$$\begin{aligned}
&\mathbf{E}[\hat{\mathbf{y}}_{t-1}^{t-1}\tilde{\mathbf{y}}_{t-2}^{t-2}] =\\
&\mathbf{E}[\{\hat{\mathbf{y}}_{t-1}^{t-2} + \mathbf{K}_{t-1}(\mathbf{B}\tilde{\mathbf{y}}_{t-1}^{t-2} + \mathbf{v}_{t-1})\}\tilde{\mathbf{y}}_{t-2}^{t-2}] =\\
&\mathbf{K}_{t-1}\mathbf{B}\mathbf{P}_{t-1,t-2}^{t-2}. && (99)
\end{aligned}$$

Finally we have,

$$\begin{aligned}
&\mathbf{E}[\hat{\mathbf{y}}_{t-1}^{t-1}\hat{\mathbf{y}}_{t-2}^{t-2}] = \mathbf{E}[(\hat{\mathbf{y}}_{t-1}^{t-2} + \mathbf{K}_{t-1}\varepsilon_{t-1})\hat{\mathbf{y}}_{t-2}^{t-2}] = \\
&\mathbf{E}[\hat{\mathbf{y}}_{t-1}^{t-2}\hat{\mathbf{y}}_{t-2}^{t-2}] = \\
&\mathbf{E}[\mathbf{y}_{t-1}\mathbf{y}_{t-2}] - \mathbf{P}_{t-1,t-2}^{t-2},
\end{aligned} \tag{100}$$

where the most important ingredient was lemma 4. Putting this together, we have

$$\begin{aligned}
&\mathbf{P}_{t-1,t-2}^{\tau} = \\
&\mathbf{P}_{t-1,t-2}^{t-2} - \mathbf{K}_{t-1}\mathbf{B}\mathbf{P}_{t-1,t-2}^{t-2} + \\
&\mathbf{J}_{t-1}\mathbf{A}\mathbf{K}_{t-1}\mathbf{B}\mathbf{P}_{t-1,t-2}^{t-2} - \mathbf{J}_{t-1}\mathbf{A}\mathbf{P}_{t-1,t-2}^{t-2}\mathbf{A}^T\mathbf{J}_{t-2}^T - \mathbf{J}_{t-1}\mathbf{A}\mathbf{Q}\mathbf{J}_{t-2}^T + \\
&\mathbf{J}_{t-1}\mathbf{P}_{t,t-1}^{\tau}\mathbf{J}_{t-2}^T
\end{aligned} \tag{101}$$

The second line can be rewriten as follows,

$$\begin{aligned}
&(I - \mathbf{K}_{t-1}\mathbf{B})\mathbf{P}_{t-1,t-2}^{t-2} = \\
&(I - \mathbf{K}_{t-1}\mathbf{B})\mathbf{P}_{t-1}^{t-2}[\mathbf{P}_{t-1}^{t-2}]^{-1}\mathbf{A}\mathbf{P}_{t-2}^{t-2} = \\
&\mathbf{P}_{t-1}^{t-1}\mathbf{J}_{t-2}^T
\end{aligned} \tag{102}$$

where we used (40) and (48). The third line can be rewritten as,

$$\begin{aligned}
&\mathbf{J}_{t-1}\mathbf{A}(\mathbf{K}_{t-1}\mathbf{B}\mathbf{P}_{t-1,t-2}^{t-2} - \mathbf{P}_{t-1,t-2}^{t-2}\mathbf{A}^T\mathbf{J}_{t-2}^T - \mathbf{Q}\mathbf{J}_{t-2}^T) = \\
&\mathbf{J}_{t-1}\mathbf{A}(\mathbf{K}_{t-1}\mathbf{B}\mathbf{P}_{t-1}^{t-2} - \mathbf{P}_{t-1,t-2}^{t-2}\mathbf{A}^T - \mathbf{Q})\mathbf{J}_{t-2}^T = \\
&\mathbf{J}_{t-1}\mathbf{A}(\mathbf{K}_{t-1}\mathbf{B}\mathbf{P}_{t-1}^{t-2} - \mathbf{P}_{t-1}^{t-2})\mathbf{J}_{t-2}^T = \\
&-\mathbf{J}_{t-1}\mathbf{A}(\mathbf{I} - \mathbf{K}_{t-1}\mathbf{B})\mathbf{P}_{t-1}^{t-2}\mathbf{J}_{t-2}^T = \\
&-\mathbf{J}_{t-1}\mathbf{A}\mathbf{P}_{t-1}^{t-1}\mathbf{J}_{t-2}^T,
\end{aligned} \tag{103}$$

where again (40) and (48) was used, and the fact that,

$$\mathbf{P}_{t-1}^{t-2} = \mathbf{P}_{t-1,t-2}^{t-2}\mathbf{A}^T + \mathbf{Q} \tag{104}$$

which can be derived analoguesly to (36). Finally, putting this together we derive the lag-one covariance smoother,

$$\mathbf{P}_{t-1,t-2}^{\tau} = \mathbf{P}_{t-1}^{t-1}\mathbf{J}_{t-2}^T + \mathbf{J}_{t-1}(\mathbf{P}_{t,t-1}^{\tau} - \mathbf{A}\mathbf{P}_{t-1}^{t-1})\mathbf{J}_{t-2}^T. \tag{105}$$