An introduction to Filtering

Samuel N. Cohen
Mathematical Institute
University of Oxford

Oxford
Mathematics

UNIVERSITY OF
OXFORD
Mathematical
Institute

Filtering

What are we doing here?

- In these lectures, we are going to look at the basic principles of 'stochastic filtering'.
- The key idea is that you have two processes, which are correlated, and you use observations of one (which you can see) to determine the behaviour of the other (which you can't see)
- Our aim is to give applicable theory, with numerical examples (all implemented in the statistical environment `R` (available at `r-project.org`))
- We will also give an example with data from high-frequency trading.

UNIVERSITY OF OXFORD
Mathematical Institute

- We begin with a simple Bayesian estimation problem, which will lead nicely to filtering.
- We have a hierarchical model for some observations $Y_1, Y_2, ..., Y_T$ with unknown mean X.
- For simplicity, suppose $X \sim N(\mu_0, \tau_0^2)$ and $Y_i|X \sim N(X, \sigma^2)$, where the Y_i are conditionally independent.
- We assume σ, τ_0, μ_0 are all known.
- Our aim is to estimate X from the observations of Y's.

We write out the joint density of X and Y_1, expand and complete the square, to see that

$$f(x,y) \propto \exp\Big(-\frac{(x-\mu_0)^2}{2\tau_0^2} - \frac{(y-x)^2}{2\sigma^2}\Big)$$
$$= \exp\Bigg(-\frac{\big(x - \frac{\mu_0/\tau_0^2 + y/\sigma^2}{1/\tau_0^2+1/\sigma^2}\big)^2}{2\frac{1}{1/\tau_0^2+1/\sigma^2}} - \frac{(y-\mu_0)^2}{2(\sigma^2+\tau_0^2)}\Bigg).$$

Using Bayes' theorem, we conclude that

$$X|Y_1 \sim N\Big(\frac{\mu_0/\tau_0^2 + Y_1/\sigma^2}{1/\tau_0^2+1/\sigma^2}, \frac{1}{1/\tau_0^2+1/\sigma^2}\Big) =: N(\mu_1, \tau_1^2).$$

- This gives us a way of 'correcting' our opinions of X given the first observation
  - We take a weighted average for the mean, and add the inverse variances ('precisions').
- Of course, we can repeat this, to include the second observation, then the third,..., and after some simplification we find

$$X|(Y_1, ..., Y_t) \sim N(\mu_t, \tau_t^2),$$

where

$$\mu_t = \frac{\mu_{t-1}/\tau_{t-1}^2 + Y_t/\sigma^2}{1/\tau_{t-1}^2 + 1/\sigma^2}, \qquad \tau_t^2 = \frac{1}{1/\tau_{t-1}^2 + 1/\sigma^2}$$

- This simplifies, in this setting, to

$$\mu_t = \frac{t\sigma^2\mu_0 + \tau_0^2\bar{Y}_t}{t\sigma^2 + \tau_0^2}, \qquad \tau_t^2 = \frac{1}{1/\tau_0^2 + t/\sigma^2}$$

  with $\bar{Y}_t = \frac{1}{t}\sum_{s=1}^{t} Y_s$.
- This simplification is special to this particular setting.

Example 1

- Let's focus on the way the distribution changes.
- Whenever we get a new observation Y_t, we correct our estimate of X, by updating the conditional distribution with the rule

$$(\mu_{t-1}, \tau_{t-1}^2) \xrightarrow{Y_t} (\mu_t, \tau_t^2) \xrightarrow{Y_{t+1}} (\mu_{t+1}, \tau_{t+1}^2).$$

- This is the basic idea of filtering: we have a hidden value X, and use our observations to update an estimate of X (in particular, its conditional distribution).

- Instead of X being constant, we will now assume that X is a random process.
- In particular, we will take $X_0 \sim N(\mu_0, \tau_0^2)$ and

$$X_t | X_{t-1} \sim N(X_{t-1}, \gamma^2), \qquad Y_t \sim N(X_t, \sigma^2).$$

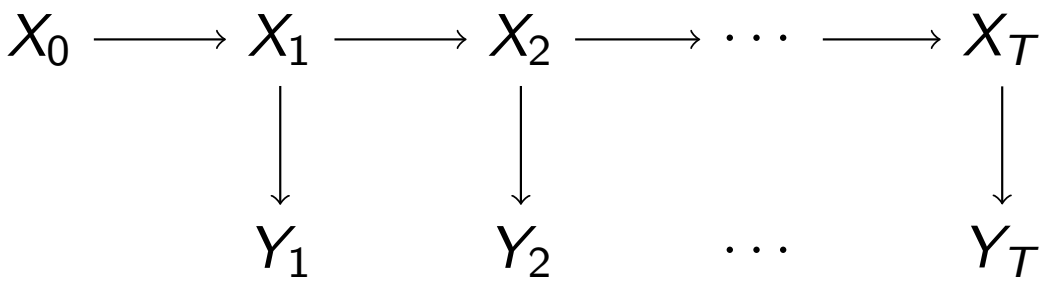
- Equivalently,

$$X_0 = \tau_0 W_0, \qquad X_t = X_{t-1} + \gamma W_t, \qquad Y_t = X_t + \sigma V_t,$$

where W, V are standard white noise (i.e. W_t, V_t are independent $N(0,1)$).

- Here γ, σ, μ_0 and τ_0 are all known.
- We call X the signal process and Y the observation process.

The dependence diagram for our model is the following:



Our conclusion depends on learning from the observations Y, but also takes account of the fact that X is changing through time.

- We want to find the distribution of X_t given $Y_1, ..., Y_t$.
- Write $\mathcal{F}_t = \sigma(X_s, Y_s; s \leq t)$ for the 'full information filtration' and $\mathcal{Y}_t = \sigma(Y_s; s \leq t)$ for the 'observation filtration'.
- We can now repeat calculations similar to those we did before:
  - From the dynamics, we have the prediction:

$$X_0 \sim N(\mu_0, \tau_0^2) \Rightarrow X_1 \sim N(\mu_0, \tau_0^2 + \gamma^2)$$

  - Writing $\tau_{1|0}^2 = \tau_0^2 + \gamma^2$, Bayes' rule gives the correction:

$$X_1|Y_1 \sim N\Big(\frac{\mu_0/\tau_{1|0}^2 + Y_1/\sigma^2}{1/\tau_{1|0}^2 + 1/\sigma^2}, \frac{1}{1/\tau_{1|0}^2 + 1/\sigma^2}\Big) =: N(\mu_{1|1}, \tau_{1|1}^2).$$

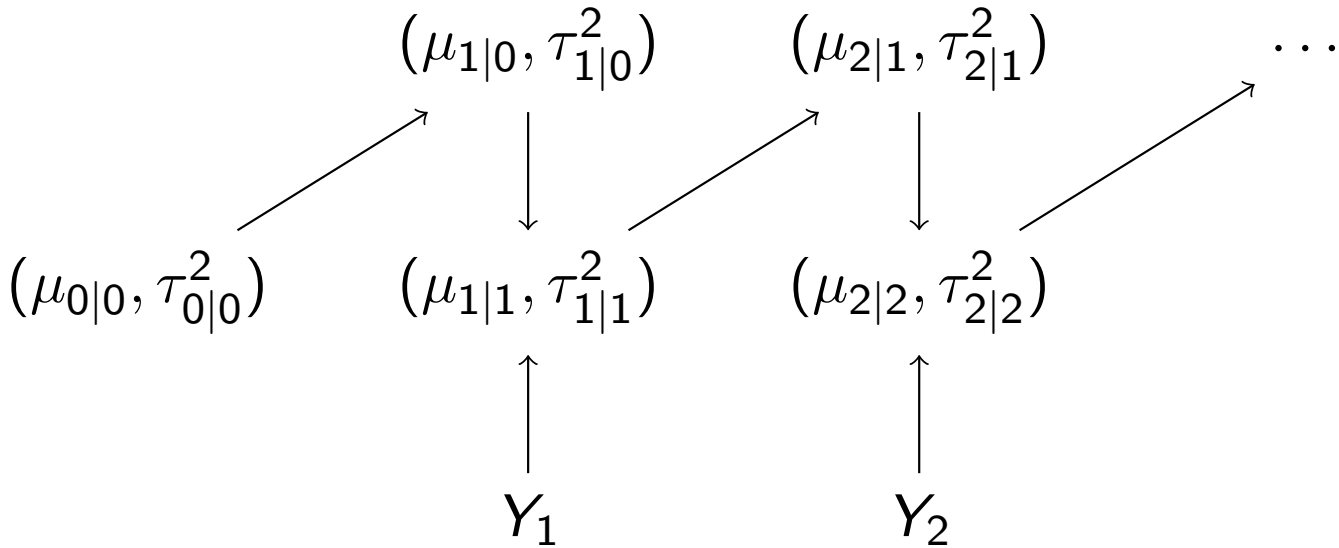
- In general, we write $\mu_{t|t-1}$ for the mean of X_t given $\mathcal{Y}_{t-1}$ and $\mu_{t|t}$ for the mean of X_t given $\mathcal{Y}_t$, similarly for the variances $\tau^2_{t|t-1}$ and $\tau^2_{t|t}$.
- Our system can be described in two steps, prediction and correction:

$$(\mu_{t-1|t-1}, \tau^2_{t-1|t-1}) \overset{\text{Prediction}}{\longrightarrow} (\mu_{t|t-1}, \tau^2_{t|t-1}) \overset{\text{Correction}}{\longrightarrow} (\mu_{t|t}, \tau^2_{t|t}),$$

where

$$\mu_{t|t-1} = \mu_{t-1|t-1}, \qquad \tau^2_{t|t-1} = \tau^2_{t-1|t-1} + \gamma^2,$$
$$\mu_{t|t} = \frac{\mu_{t-1|t-1}/\tau^2_{t|t-1} + Y_t/\sigma^2}{1/\tau^2_{t|t-1} + 1/\sigma^2}, \qquad \tau^2_{t|t} = \frac{1}{1/\tau^2_{t|t-1} + 1/\sigma^2}.$$

- By iterating these equations, we solve our filtering problem, that is, we have a complete description of the distribution of X_t given $\mathcal{Y}_t$ for every t.
- These calculations are recursive, so including new observations is simple (and fast!).



Example 2

- These equations can be solved quickly, using only basic methods.
- Updating only involves addition, multiplication and division.
  - Division can be largely avoided by using precisions (τ^{-2}) instead of variances (τ^2).
- Observe that $\tau^2_{t|t}$ converges quickly to a stationary value, the limit is found by solving the equation

$$\tau^2 = \frac{1}{1/(\tau^2+\gamma^2)+1/\sigma^2} \Rightarrow \tau^2 = \frac{1}{2}\left(\gamma\sqrt{4\sigma^2+\gamma^2}-\gamma^2\right).$$

- We have just solved a simple case of the famous Kalman filtering problem.
- The general case has two differences: our processes are vector valued and the relationship between X and Y is more general (but still linear).
- These simple generalizations yield an extraordinarily powerful technique.

Consider the following model:

$$X_t = AX_{t-1} + W_t, \qquad Y_t = CX_t + V_t$$

with starting distribution $X_0 \sim N(\mu_{0|0}, P_{0|0})$.
Here

- W, V are white noise processes in $\mathbb{R}^k$ and $\mathbb{R}^d$ with respective (nonnegative definite) variances Γ and Σ
  - In other words, $W_t \sim N(0, \Gamma)$ and $V_t \sim N(0, \Sigma)$ for all t, all values independent.
- A and Γ are $k \times k$-matrices, C is $d \times k$, Σ is $d \times d$.
- We know A, C, Γ, Σ.

The key fact we will need is that

> *if you have jointly (multivariate) normal random variables Y, Z, then $Y|Z$ is also normal.*

Furthermore

$$E[Y|Z] = E[Y] + \mathrm{cov}(Y,Z)\mathrm{var}(Z)^{-1}(Z - E[Z])$$
$$\mathrm{var}(Y|Z) = \mathrm{var}(Y) - \mathrm{cov}(Y,Z)\mathrm{var}(Z)^{-1}\mathrm{cov}(Y,Z)^{\top}$$

These facts can be proven using the densities, and justify everything that follows.

The Kalman Filter: Prediction

Step 1 of the filter

UNIVERSITY OF OXFORD
Mathematical Institute

- We know that $X_t|\mathcal{Y}_t = X_t|(Y_1, Y_2, ..., Y_t)$ is normal (and similarly $X_t|\mathcal{Y}_{t-1}$),
- Using the dynamics of X and Y, we can easily calculate the prediction equations:

$$\begin{aligned}
\mu_{t|t-1} &= E[X_t|\mathcal{Y}_{t-1}] = E[AX_{t-1} + W_t|\mathcal{Y}_{t-1}] \\
&= AE[X_{t-1}|\mathcal{Y}_{t-1}] \\
&= A\mu_{t-1|t-1} \\
P_{t|t-1} &= \mathrm{var}(X_t|\mathcal{Y}_{t-1}) = \mathrm{var}(AX_{t-1} + W_t|\mathcal{Y}_{t-1}) \\
&= A\mathrm{var}(X_{t-1}|\mathcal{Y}_{t-1})A^\top + \mathrm{var}(W_t|\mathcal{Y}_{t-1}) \\
&= AP_{t-1|t-1}A^\top + \Gamma
\end{aligned}$$

- The correction equations are made simpler if we first calculate the 'innovation' process η and its variance S
- η tells us what 'new' information we learn from Y_t

$$\eta_t = Y_t - E[Y_t|\mathcal{Y}_{t-1}] = Y_t - C\mu_{t|t-1},$$
$$S_t = \mathrm{var}(\eta_t|\mathcal{Y}_{t-1}) = \mathrm{var}(Y_t|\mathcal{Y}_{t-1}) = CP_{t|t-1}C^\top + \Sigma.$$

- Using S, we can calculate the 'Kalman gain' process, which allows us to optimally incorporate new information,

$$K_t = P_{t|t-1}C^\top S_t^{-1} = (S_t^{-1}CP_{t|t-1})^\top$$

Finally, it is easy to calculate the correction equations:

$$\mu_{t|t} = \mu_{t|t-1} + K_t \eta_t,$$
$$P_{t|t} = (I - K_t C) P_{t|t-1}.$$

Given these equations, we are ready to calculate!

Example 3

- Using our equations, it is easy to see how to calculate the forecasted values $E[X_t|\mathcal{Y}_s]$ for $s < t$.
- By direct recursion:

$$\mu_{t|s} = E[X_t|\mathcal{Y}_s] = A^{t-s}\mu_{s|s}.$$

- Furthermore, the conditional variance $P_{t|s} = \mathrm{var}(X_t|\mathcal{Y}_s)$ satisfies

$$P_{t+1|s} = AP_{t|s}A^\top + \Gamma$$

which is easy to calculate recursively.

The Kalman Filter: Smoothing

Harder, but useful!

UNIVERSITY OF OXFORD
Mathematical Institute

- Calculating the 'smoother', that is, $\mu_{t|N} = E[X_t|\mathcal{Y}_N]$ for $t < N$ is also possible.
- First write $J_t = P_{t|t}A^\top P^{-1}_{t+1|t}$. Then, using our basic properties of normal distributions (and plenty of algebra),

$$\mu_{t|N} = \mu_{t|t} + J_t(\mu_{t+1|N} - \mu_{t+1|t}),$$
$$P_{t|N} = P_{t|t} + J_t(P_{t+1|N} - P_{t+1|t})J_t^\top,$$

- These can be calculated backwards, starting at time N.
- In effect, you first do a single forward pass through the observations from $0 \to N$ calculating the filter, then backwards from $N \to 0$ to calculate the smoother.

Example 3 (ctd)

- We shall see that, when trying to fit a filter in practice, it will also be useful to know the values of

$$P_{t-1,t|N} := E[(X_t - \mu_{t|N})(X_{t-1} - \mu_{t-1|N})^\top | \mathcal{Y}_N].$$

- Fortunately, there is a formula:

$$P_{N-1,N|N} = (I - K_N C)AP_{N-1|N-1},$$
$$P_{t-1,t|N} = P_{t|t}J_{t-1}^\top + J_t(P_{t,t+1|N} - AP_{t|t})J_{t-1}^\top.$$

- The derivation is even more algebra than before.
- It can also be calculated using a single sweep back through the data.

Exercise: prove these formulae!

To see how rich a theory this gives, consider an ARMA(1,1) process, where

$$x_t = \phi x_{t-1} + \theta z_{t-1} + z_t$$

for constants ϕ, θ and white noise z.

- We only observe x_t.
- It's difficult to calculate $E[x_t | x_{t-1}, x_{t-2}, ...]$, which is usually needed when fitting these models.
- This does not look like the models we've considered...

To see how rich a theory this gives, consider an ARMA(1,1) process, where

$$x_t = \phi x_{t-1} + \theta z_{t-1} + z_t$$

for constants ϕ, θ and white noise z.

- We only observe x_t.
- It's difficult to calculate $E[x_t | x_{t-1}, x_{t-2}, ...]$, which is usually needed when fitting these models.
- This does not look like the models we've considered... until we write it as a 'state space' model.

Example: An ARMA(1,1) process

Surprisingly a Kalman Filter model!

Mathematical Institute

We can write

$$X_t = \begin{bmatrix} x_t \\ \theta z_t \end{bmatrix} = \begin{bmatrix} \phi & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ \theta z_{t-1} \end{bmatrix} + \begin{bmatrix} 1 \\ \theta \end{bmatrix} z_t = \begin{bmatrix} \phi & 1 \\ 0 & 0 \end{bmatrix} X_{t-1} + W_t$$

and

$$Y_t = x_t = \begin{bmatrix} 1 & 0 \end{bmatrix} X_{t-1}.$$

- Hence we can apply the Kalman filter to X, and so efficiently calculate

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \mu_{t|t-1} = \begin{bmatrix} 1 & 0 \end{bmatrix} E[X_t|\mathcal{Y}_{t-1}] = E[x_t|x_{t-1}, x_{t-2}, \ldots].$$

- In our earlier notation, we have $\Sigma = 0$, A, C as indicated and

$$\Gamma = \begin{bmatrix} 1 & \theta \\ \theta & \theta^2 \end{bmatrix}.$$

- The equations we've seen have been fairly 'nice'.
- The filters can be solved in closed form, recursively, and are finite dimensional.
- This is because we have assumed throughout that all our random variables are Gaussian, and all the relationships between them are linear.
- Without this assumption, as we will see in continuous time, we are in a much more difficult situation.
- One other case where a nice set of equations can be obtained is when X is a finite-state Markov chain.

- Suppose X is a finite-state Markov chain. We write X as a process $X_t = AX_{t-1} + M_t$ where X takes values in the basis vectors in $\mathbb{R}^d$, and M is a martingale difference process (so $E[M_t|\mathcal{F}_{t-1}] = 0$).
- The matrix $A^\top$ is the familiar transition matrix of the Markov chain.
- We just need to calculate the probability X takes values in each state, or equivalently, the vector $\mu_{t|t} = E[X_t|\mathcal{Y}_t] \in \mathbb{R}^d$ (as $P(X_t = e_i|\mathcal{Y}_t) = E[e_i^\top X_t|\mathcal{Y}_t] = e_i^\top \mu_{t|t}$
- We assume that $Y_t|\mathcal{F}_t \sim c(y; X_t)m(dy)$, where c is some density function and m is some measure (no normality is needed).

- We can directly calculate the prediction equation:

$$\mu_{t|t-1} = E[X_t|\mathcal{Y}_{t-1}] = E[AX_{t-1} + M_t|\mathcal{Y}_{t-1}] = A\mu_{t-1|t-1}.$$

- To calculate the correction equation, we use Bayes' theorem:

$$\begin{aligned} P(X_t = e_i|Y_t, \mathcal{Y}_{t-1}) &= \frac{c(Y_t; e_i)P(X_t = e_i|\mathcal{Y}_{t-1})}{\sum_j c(Y_t; e_j)P(X_t = e_j|\mathcal{Y}_{t-1})} \\ &\propto c(Y_t; e_i)P(X_t = e_i|\mathcal{Y}_{t-1}) \end{aligned}$$

- Again, forecasting is easy: $\mu_{t|s} = A^{t-s}\mu_{s|s}$ for $s < t$.
- Smoothing can be done with a backward pass, by looking at a 'dual' variable ν satisfying the equation (for $N > t$)

$$\nu_{t|N} \propto A^\top C(Y_{t+1})\nu_{t+1|N}, \qquad \nu_{N|N} = 1,$$

  and then calculating $\mu_{t|N} \propto \mu_{t|t}\nu_{t|N}$, where the product is taken component by component.
- There are closed form equations for other quantities also (for example, estimating occupation times, the number of transitions, functions of X and Y, ... see Elliott, Aggoun and Moore, Hidden Markov Models, Springer 1995)

Example 4

- So now we move gear a little technically, as we want to see what happens in continuous time.
- This is particuarly useful as a model when observations occur in very high frequency, as it allows us to find good approximations to our problem.
- On the other hand, it becomes more difficult to find and solve the filtering equations.

- The approach we shall take is called the 'reference probability method'.
- It depends on the following result, which will serve as "Bayes' theorem" in this context.

Theorem

Suppose we have a probability measure $Q \sim P$. Write the Radon–Nikodym density $Z = dQ/dP$, and suppose we have a filtration $\{\mathcal{F}_t\}_{t\geq 0}$. Then for any $t \geq 0$ and any random variable ξ, we know that

$$E_Q[\xi|\mathcal{F}_t] = \frac{E_P[Z\xi|\mathcal{F}_t]}{E_P[Z|\mathcal{F}_t]}.$$

- We assume as before that we have processes X and Y, on an interval $[0, T]$.
- These satisfy the SDEs

$$dX_t = f(t, X_t)dt + \kappa(t, X_t)dB_t$$
$$dY_t = c(t, X_t)dt + dW_t$$

  where f, κ, c are known (Lipschitz continuous) functions, and B and W are Brownian motions.
- We assume X and Y are scalar and B and W are independent for simplicity.
  - These assumptions can be relaxed, but the notation becomes more difficult.

- From the Feynman–Kac theorem/Ito's lemma, we know that for any smooth bounded function ϕ,

$$\phi(X_t) = \phi(X_0) + \int_{[0,t]} \mathcal{L}\phi(X_u)du + \text{martingale}$$

where $\mathcal{L}$ is the infinitesimal generator of X, that is,

$$\mathcal{L}\phi = \frac{\partial \phi}{\partial x} f(t,x) + \frac{1}{2} \cdot \frac{\partial^2 \phi}{\partial x^2} \kappa(t,x)^2.$$

- We expect $\mathcal{L}$ to be part of the solution to our filtering problem.

- We define a probability Q by $\frac{dQ}{dP} = Z_T$, where

$$Z_t = \mathcal{E}\Big(-\int c(s, X_s) dW_s\Big)_t = \exp\Big(-\int_0^t c(s, X_s) dW_s - \frac{1}{2}\int_0^t c(s, X_s)^2 dt\Big).$$

- We write $\Lambda = 1/Z$, and using Ito's lemma we can see that $d\Lambda_t = \Lambda_t c(t, X_s) dY_s$.
- Using Girsanov's theorem, this change of measure has the effect of changing the drift in Y, so under Q we have the dynamics

$$dX_t = f(t, X_t)dt + \kappa(t, X_t)dB_t, \qquad dY_t = dW_t^Q$$

where B and W^Q are independent Q-Brownian motions.

- X and Y are independent under Q!

- We will now try to calculate the *unnormalized* expectations, which we write:

$$\sigma_t(\phi) := E_Q[\Lambda_t \phi(X_t)|\mathcal{Y}_t].$$

- "Bayes' theorem" tells us that $E_P[\phi(X_t)|\mathcal{F}_t] = \sigma_t(\phi)/\sigma_t(1)$.
- Now, we can write out $\Lambda_s \phi(X_s)$ using Ito's lemma. This gives

$$\begin{aligned} d(\Lambda\phi(X))_t &= \Lambda_t \frac{\partial \phi}{\partial x} dX_t + \frac{1}{2}\Lambda_t \frac{\partial^2 \phi}{\partial x^2}\kappa(t, X_t)^2 dt + \Lambda_t \phi(X_t) c(t, X_t) dY_t \\ &= \Lambda_t \mathcal{L}\phi(X_t) dt + \Lambda_t \frac{\partial \phi}{\partial x}\kappa(t, X_t) dB_t + \Lambda_t \phi(X_t) c(t, X_t) dY_t \end{aligned}$$

- Taking an expectation, as (X, B) and Y are Q-independent, we have the 'Zakai equation'

$$\begin{aligned}\sigma_t(\phi) &= E_Q[\Lambda_t\phi(X_t)|\mathcal{Y}_t] \\ &= \sigma_0(\phi) + \int_0^t E_Q[\Lambda_s\mathcal{L}\phi(X_s)|\mathcal{Y}_t]ds + \int_0^t E_Q[\Lambda_s\phi(X_s)c(s,X_s)|\mathcal{Y}_t]dY_s \\ &= \sigma_0(\phi) + \int_0^t E_Q[\Lambda_s\mathcal{L}\phi(X_s)|\mathcal{Y}_s]ds + \int_0^t E_Q[\Lambda_s\phi(X_s)c(s,X_s)|\mathcal{Y}_s]dY_s \\ &= \sigma_0(\phi) + \int_0^t \sigma_s(\mathcal{L}\phi)ds + \int_0^t \sigma_s(\phi c)dY_s\end{aligned}$$

- This is a simple equation apart from one thing: the term $\sigma_s(\phi c)$ cannot be calculated recursively in terms of $\sigma_s(\phi)$.

- Rearranging and applying Ito's lemma, we can obtain an equation for the *normalized* expectations

$$\pi_s(\phi) := \sigma_s(\phi)/\sigma_s(1) = E[\phi(X_s)|\mathcal{Y}_s],$$

the 'Fujisaki–Kallianpur–Kunita' equation

$$\pi_t(\phi) = \pi_0(\phi) + \int_{[0,t]} \pi_s(\mathcal{L}\phi)du + \int_{[0,t]} \big(\pi_s(\phi c) - \pi_s(\phi)\pi_s(c)\big)dV_s.$$

- Here $dV_s = dY_s - \pi_s(c)ds$ is the (differential of the) 'innovations process' (and is a $\mathcal{Y}$-Brownian motion under P).

- Let's assume X has a smooth density given $\mathcal{Y}_t$, so $\sigma_t(\phi) = \int_{\mathbb{R}} \phi(x) q(t,x) dx$. Then we see that

$$\int_{\mathbb{R}} \phi(x) q(t,x) dx = \int_{\mathbb{R}} \phi(x) q(0,x) dx + \int_0^t \int_{\mathbb{R}} \mathcal{L}\phi(x) q(s,x) dx ds + \int_0^t \int_{\mathbb{R}} \phi(x) c(s,x) q(s,x) dx \, dY_s$$

- By integration by parts, if $\mathcal{L}^*$ is the adjoint of $\mathcal{L}$

$$\mathcal{L}^* q = \frac{\partial (qf)}{\partial x} + \frac{1}{2} \cdot \frac{\partial^2 (q\kappa)}{\partial x^2},$$

we calculate

$$\int_{\mathbb{R}} \phi(x) q(t,x) dx = \int_{\mathbb{R}} \phi(x) \Big(q(0,x) + \int_0^t \mathcal{L}^* q(s,x) ds + \int_0^t c(s,x) q(s,x) dY_s \Big) dx$$

- This should hold for every smooth and bounded ϕ, so we have the linear SPDE

$$q(t,x) = q(0,x) + \int_0^t \mathcal{L}^* q(s,x) ds + \int_0^t c(s,x) q(s,x) dY_s$$

We can then calculate the density of $X_t | \mathcal{Y}_t$ as

$$p(t,x) = \frac{q(t,x)}{\int_{\mathbb{R}} q(t,x') dx'}.$$

- One can also get a nonlinear SPDE for the normalized density.
- Solving SPDEs is hard, so this equation is not frequently solved in practice in this general form – instead it suggests good approximations, or allows special cases to be derived.

- Let's see the continuous-time Gaussian case.
- Here we assume $c(t, X_t) = cX_t$, $f(t, X_t) = aX_t$ and $\kappa(t, X_t) = b$. Then we have the dynamics

$$dX_t = aX_t dt + b\, dB_t, \qquad dY_t = cX_t dt + dW_t$$

- Here a, b, c are known.
- With $\tilde{Y} = Y/c$ and $f = 1/c$, this is the same as the model for observations $d\tilde{Y}_t = X_t dt + f\, dW_t$.
- We know that these equations define a Gaussian process (i.e. all marginals are jointly normal), so it's enough to calculate the mean and variance.

- Write $\hat{X}_t = E_P[X_t|\mathcal{Y}_t]$.
- First observe that everything here is Gaussian, and $\hat{X} - X$ is uncorrelated with Y_s for all $s < t$.
- In particular, this implies they are independent, and

$$E[(X_t - \hat{X}_t)^2|\mathcal{Y}_t] = E[(X_t - \hat{X}_t)^2] =: P_t$$

  is deterministic.
- Also, $E[(X_t - \hat{X}_t)^3|\mathcal{Y}_t] = 0$.

- Taking $\phi(x) = x$ so $\hat{X}_t = \pi_t(\phi)$, we know $\mathcal{L}\phi \equiv 0$, so

$$\begin{aligned}\hat{X}_t &= \hat{X}_0 + \int_0^t a\hat{X}_s ds + c\int_0^t \left(\pi_s(X_s^2) - \hat{X}_s^2\right) dV_s \\ &= \hat{X}_0 + \int_0^t a\hat{X}_s ds + c\int_0^t P_s dV_s.\end{aligned}$$

- Notice this is in terms of the innovations process V.

- Taking $\phi(x) = x^2$,

$$\pi_t(X^2) = \pi_0(X^2) + \int_{[0,t]} (2a\pi_s(X^2) + b^2)du$$
$$+ c \int_{[0,t]} (\pi_s(X^3) - \hat{X}_u \pi_u(X^2))dV.$$
$$\hat{X}_t^2 = \hat{X}_0^2 + \int_0^t 2a(\hat{X}_s)^2 ds + 2c \int_0^t \hat{X}_s P_s dV_s.$$

- Taking a difference and simplifying, we obtain a Riccatti equation for the variance P

$$P_t = \pi_t(X^2) - \hat{X}_t^2 = P_0 + \int_0^t (2aP_s + b^2 - c^2 P_s^2)du.$$

- Together, we have an SDE for the mean

$$\hat{X}_t = \hat{X}_0 + \int_0^t a\hat{X}_s ds + c\int_0^t P_s dV_s.$$

and a (deterministic) Riccatti equation for the variance

$$P_t = P_0 + \int_0^t (2aP_s + b^2 - c^2 P_s^2) du.$$

- This pair of equations is called the 'Kalman–Bucy filter'.
- It can then be approximated using the usual methods for SDEs/ODEs (eg Euler methods)
- It is possible to obtain a Kalman–Bucy smoother as well.

Example 5

- Just as in discrete time, there is a continuous time equation for the filter based on a (continuous time) Markov chain.
- Here we have the dynamics

$$dX_t = AX_t dt + dM_t$$
$$dY_t = c^\top X_t dt + dW_t$$

where $A^\top$ is the Q-matrix of the Markov chain, M is a martingale and c is a vector.

- As X is written using only basis vectors in $\mathbb{R}^N$, any function of X can be written $\Phi(X) = \phi^\top X$ for some vector $\phi \in \mathbb{R}^N$.
- While X is not of the form we considered earlier, we can still find the generator of X is $\mathcal{L}\Phi = A^\top \phi$, and the adjoint of $\mathcal{L}$ is simply $\mathcal{L}^* v = Av$.
- We can calculate (from the Zakai equation) the unnormalized probability vector for the state of X

$$E[X_t|\mathcal{Y}_t] \propto q_t = q_0 + \int_0^t A q_u du + \int_0^t \mathrm{diag}(c) q_u dY_u.$$

- This equation is just an N-dimensional linear SDE. Equations for the smoother are also known.

Calibrating a filter

Trying to make things useable.

Mathematical Institute

- What we have seen so far deals with the problem of how to take our observations Y and obtain the behaviour of X.
- However, we have assumed throughout that we know the probability model, that is, all the other parameters are fixed.
- The question of how to estimate those parameters is what we consider next.
- This problem has a wide range of approaches, depending on the details involved.
- We shall focus on a simple case, using the EM algorithm (discussed in Elliott, van Der Hoek and Malcolm (2005), based on Shumway and Stoffer (1982)).

- We focus on the following simple scalar version of the discrete time Kalman filter:

$$X_{t+1} = a + bX_t + cW_{t+1}$$
$$Y_t = X_t + f\,V_t.$$

- If we could observe X and Y directly, then we could calculate a, b, c and f easily using regression.
- If we cannot observe X, then we need to use a more advanced method.

- Our filtering equations simplify to

$$\mu_{t+1|t} = a + b\mu_{t|t}, \qquad P_{t+1|t} = b^2 P_{t|t} + c^2$$

$$K_{t+1} = \frac{P_{t+1|t}}{P_{t+1|t} + f^2}$$

$$\mu_{t+1|t+1} = \mu_{t+1|t} + K_{t+1}(y_{t+1} - \mu_{t+1|t})$$

$$P_{t+1|t+1} = (1 - K_{t+1})P_{t+1|t} = f^2 K_{t+1}$$

- The smoothing equations are (with $J_t = bP_{t|t}/P_{t+1|t}$)

$$\mu_{t|N} = \mu_{t|t} + J_t\big(\mu_{t+1|N} - (a + b\mu_{t|t})\big)$$

$$P_{t|N} = P_{t|t} + J_t^2\big(P_{t+1|N} - P_{t+1|t}\big)$$

$$P_{t-1,t|N} = J_{t-1}P_{t|t} + J_t J_{t-1}\big(P_{t,t+1|N} - bP_{t|t}\big)$$

$$P_{N-1,N|N} = b(1 - K_N)P_{N-1|N-1}$$

- So, how to estimate the parameters?
- Simply regressing the smoothed values of X gives bias, as we expect X will be 'rougher' than the smoothed values.
- The likelihood function is hard to compute, as it depends on X and Y
- If we assumed we could calculate the expectation, then we could instead try to maximize the expected log-likelihood $E[\ell(a, b, c, f; \{X_t, Y_t\}_{t \leq T}) | \mathcal{Y}_T]$
- We can then iterate (calculate parameters) $\leftrightarrow$ (calculate filter estimates) until convergence. This is the "Expectation-Maximization algorithm", as we iterate between (Maximum likelihood step) $\leftrightarrow$ (Expectation step).

- The maximization step can be solved! (all sums from 1 to N):

$$\hat{b} = \frac{E[\sum(X_{t-1} - \frac{1}{N}\sum X_{t-1})(X_t - \frac{1}{N}\sum X_{t-1})|\mathcal{Y}_N]}{E[\sum(X_t - \bar{X})^2|\mathcal{Y}_N]}$$

$$= \frac{\sum P_{t-1,t|N} + \sum \mu_{t|N}\mu_{t-1|N} - \frac{1}{N}\sum \mu_{t|N}\sum \mu_{t-1|N}}{\sum P_{t|N} + \sum \mu_{t|N}^2 - \frac{1}{N}(\sum \mu_{t|N})^2}$$

$$\hat{a} = \frac{1}{N}\sum \mu_{t|N} - \hat{b}\frac{1}{N}\sum \mu_{t-1|N}$$

$$\hat{c} = \frac{1}{N}E\left[\sum(X_t - \hat{a} - \hat{b}X_{t-1})^2 \middle| \mathcal{Y}_N\right]$$

$$= \frac{1}{N}\sum \Big(P_{t|N} + \mu_{t|N}^2 + \hat{a}^2 - 2\hat{a}\mu_{t|N} + 2\hat{a}\hat{b}\mu_{t-1|N}$$

$$+ \hat{b}^2(P_{t-1|N} + \mu_{t-1|N}^2) - 2\hat{b}P_{t-1,t|N} - 2\hat{b}\mu_{t|N}\mu_{t-1|N}\Big)$$

$$\hat{f} = \frac{1}{N}\sum \Big((Y_t - \mu_{t|N})^2 + P_{t|N}\Big)$$

- We start off with approximate estimates of a, b, c, f
- We iterate the EM algorithm to improve these estimates
- Convergence may be slow (or get stuck) given bad starting points.
- Given a large amount of data, it may be worth starting with only a small subsample, then increasing the amount of data used as you go.

Example 6

- We will look at using these methods to creat a basic pairs trading system, using a toy setup, with real data.
- We will build this using one-second mid-prices for Microsoft (MSFT) and Intel Corp. (INTC), on individual days in the week beginning 3 November 2014.
- Thanks to Álvaro Cartea and Sebastian Jaimungal for data.

- We will model this using the method in previous section, as suggested by Elliott, van der Hoek and Malcolm (2005).
- We fit the filter using $Y = \log(\texttt{INTC}/\texttt{MSFT})$ using the Kalman–EM method described above.
- We will then create a trading signal depending on the value of $Y_t - \mu_{t|t}$.
- If our model is reasonable, we expect this value will revert quickly to zero, which suggests a profitable trade, either long INTC and short MSFT (if $Y < \mu_{t|t}$) or vice versa.
- Effectively, we expect prices to oscillate around a short-term mean

- We choose to trade only when the difference is sufficiently large, in such a way that we have a position 1% of the time.
- We reevaluate our position every second, and only invest/short at most \$1 in each stock.
- We ignore all transaction costs, microstructure issues, trading constraints, etc.

Example 7

- This suggests that these methods can be used to build profitable trading strategies.
- Of course, we would need to incorporate further effects into our model of profits before implementing this in practice, as in the real world we can only buy at the ask and sell at the bid, which will likely eliminate most of our observed gains.
- Filtering is fast, which is important in this setting.

- We have looked at the problem of filtering in a variety of contexts.
  - Discrete/Continuous time
  - Gaussian/Finite state (or general with an SPDE solution)
- We have seen how you can implement these filters, and how to estimate the coefficients in a simple setting
- The EM algorithm can be used more widely
- We have seen a toy application of these methods to financial data