Topic 4: Filtering methods and learning

Jaroslav Borovička

Computational Dynamics (Spring 2023)

New York University

Content

Economic problem

- In a variety of setups, agents do not have perfect information about the environment.
- They however use available information to update their beliefs about the state of the economy.
- How exactly is this information processed? Bayesian agents use information efficiently but there are other belief updating schemes as well.

Tools

- Kalman filter
- Relationship to VAR estimation

Literature

Textbook

- Ljungqvist and Sargent (2020), Chapter 2 (Sections 2.7–2.9, Kalman filter)

QuantEcon

- Quantitative Economics with Python: Topic 31 (Kalman filter), Topic 50–52 (Bayes law), Topic 53 (Search with learning)

Kalman filter

Jovanovic (1979) consider the problem of a worker-firm pair that learns about the productivity of the worker in the match.

- worker in the given firm has true productivity θ
- the worker-firm pair starts with some prior belief about this true productivity parameter
- every period, they observe worker's output, which can be interpreted as a noisy realization of θ

$$y_t = \theta + v_t$$

- given this observation, they update their belief about θ

How should such belief updating be conducted?

- Bayes law!
- the key question is how to keep the problem tractable
- the belief is a whole distribution over the support of θ

We consider the linear model

$$\begin{array}{rcll} x_{t+1} & = & A_o x_t + C w_{t+1} & n \times 1 \qquad w_{t+1} \sim N(0, I_p) \\ y_t & = & G x_t + v_t & m \times 1 \qquad v_t \sim N(0, R)\,. \end{array} \tag{4.1}$$

- x_t is an unobservable vector with a law of motion given by a Gaussian VAR
- y_t is a measurement vector
- v_t is measurement noise independent of $\{w_t\}_{t=1}^{\infty}$
- we assume an initial condition

$$x_0 \sim N(\hat{x}_0, \Sigma_0)\,. \tag{4.2}$$

The problem is to construct the optimal forecast of the path of the underlying state, given the observed measurements.

- solved by Kalman (1960)
- the algorithm leads to a recursive formula for the best predictor of x_t given observations $y^{t-1} = (y_{t-1}, \ldots, y_0)$
- because of the linear-Gaussian structure of the whole system, we can infer that the predictor will also be Gaussian, so it is sufficient to construct predictors for the first two moments of its distribution

$$\begin{aligned} \hat{x}_t &= E\left[x_t \mid y^{t-1}\right] \\ \Sigma_t &= E\left[(x_t - \hat{x}_t)(x_t - \hat{x}_t)'\right]. \end{aligned} \tag{4.3}$$

- this is a critical simplification of the problem because we do not need to keep track of the whole distribution in the form of a function over the state space

Since the underlying state x_t is persistent, observed data y_t will be serially correlated.

- The idea is to derive the contribution of the information embedded in the new observation y_t to the prediction of x_{t+1}, relative to what we already could have inferred from y^{t-1}.
- Mathematically, we are applying the **Gram–Schmidt orthogonalization process** to the sequence of the data observations y^t, and constructing an orthogonal basis of **innovations** $(a_0, a_1, \dots a_t)$ from $(y_0, y_1, \dots y_t)$.
- A particular innovation a_t will represent new information in the additional data point y_t relative to what we learned from y^{t-1}. Since the innovations are orthogonal, projecting the unobserved state x_t on the innovations will be simpler than projecting on the original data.
- Effectively, we are replacing (4.3) with an equivalent representation

$$\hat{x}_t = E\left[x_t \mid a^{t-1}\right].$$

Let us start from the initial condition for the prior $x_0 \sim N(\hat{x}_0, \Sigma_0)$.

- this prior represents the agent's belief before any data have been observed
- we now study how much we can learn about the unknown x_0 from observing y_0

We will proceed by using linear projections, which corresponds to running theoretical OLS regressions.

- this method is justified as a way of obtaining efficient forecasts given the linear-Gaussian environment for the problem

Since

$$y_t = Gx_t + v_t$$

we infer that the prior belief about y_0 before y_0 has been observed is

$$x_0 \sim N(\hat{x}_0, \Sigma_0) \qquad \Longrightarrow \qquad y_0 \sim N\left(G\hat{x}_0, G\Sigma_0 G' + R\right).$$

The agent expects y_0 to be on average $G\hat{x}_0$, so $y_0 - G\hat{x}_0$ can be viewed as a surprise, or innovation, relative to the expected value.

To construct the forecast of x_0 given the observation of the data point y_0

- we can split $x_0 = \hat{x}_0 + (x_0 - \hat{x}_0)$, where $\hat{x}_0$ is known and $x_0 - \hat{x}_0$ is the unknown part
- similarly, we can split $y_0 = G\hat{x}_0 + (y_0 - G\hat{x}_0)$

Let us project the unknown $x_0 - \hat{x}_0$ on the new information embedded in the observation of y_0:

$$x_0 - \hat{x}_0 = L_0 \underbrace{(y_0 - G\hat{x}_0)}_{\text{innovation } a_0} + \eta_0$$

- the residual η_0 is orthogonal on the innovation $a_0 \doteq y_0 - G\hat{x}_0$ by construction
- the innovation a_0 represents the 'surprise', or new information, embedded in the observation of y_0 relative to its expected value $G\hat{x}_0$
- a_0 contains the same information as y_0
- this innovation also constitutes the first element in the construction of the orthogonal basis $(a_0, a_1, \ldots)$ constructed from observations $(y_0, y_1, \ldots)$.

Given the orthogonality between η_0 and a_0, post-multiplying by $(y_0 - G\hat{x}_0)'$ and taking expectations yields

$$\underbrace{E\left[(x_0 - \hat{x}_0)(y_0 - G\hat{x}_0)'\right]}_{= Cov(x_0, y_0|\hat{x}_0, \Sigma_0)} = L_0 \underbrace{E\left[(y_0 - G\hat{x}_0)(y_0 - G\hat{x}_0)'\right]}_{= Var(y_0|\hat{x}_0, \Sigma_0)} + \underbrace{E\left[\eta_0 (y_0 - G\hat{x}_0)'\right]}_{= 0}$$

so that

$$\Sigma_0 G' = L_0 \left(G\Sigma_0 G' + R\right).$$

and hence the $n \times m$ matrix regression coefficient L_0 takes the form

$$L_0 = \Sigma_0 G' \left(G\Sigma_0 G' + R\right)^{-1}.$$

Equation (4.1) then implies that we can write

$$x_1 = A_o x_0 + C w_1 = A_o \hat{x}_0 + A_o \left(x_0 - \hat{x}_0\right) + C w_1 . \tag{4.4}$$

- $A_o \hat{x}_0$ is the best prediction of x_1 based on the prior
- $A_o \left(x_0 - \hat{x}_0\right)$ represents uncertainty about x_1 inherited from prior uncertainty about x_0
- $C w_1$ is the new uncertainty linked to the new random innnovation

The mean forecast of the state x_1 given the data point y_0 is therefore given by

$$\begin{aligned} \hat{x}_1 &= E\left[x_1 | y^0\right] = E\left[x_1 | a^0\right] = A_o \hat{x}_0 + E\left[A_o \left(x_0 - \hat{x}_0\right) | a^0\right] \\ &= A_o \hat{x}_0 + \underbrace{A_o L_0}_{= K_0} \left(y_0 - G \hat{x}_0\right), \end{aligned} \tag{4.5}$$

where the matrix

$$K_0 = A_o \Sigma_0 G' \left(G \Sigma_0 G' + R\right)^{-1}$$

is called the **Kalman gain**.

- the Kalman gain captures how informative the new innovation is for predicting x_1

Subtracting (4.5) from (4.4) yields

$$\begin{aligned} x_1 - \hat{x}_1 &= A_o (x_0 - \hat{x}_0) + Cw_1 - K_0 (y_0 - G\hat{x}_0) \\ &= (A_o - K_0 G)(x_0 - \hat{x}_0) + Cw_1 - K_0 v_0. \end{aligned}$$

Notice that the three terms on the previous line are independent. Hence the variance of the forecast given the data point y^0

$$\begin{aligned} \Sigma_1 &= E\left[(x_1 - \hat{x}_1)(x_1 - \hat{x}_1)'\right] = \\ &= (A_o - K_0 G)\Sigma_0 (A_o - K_0 G)' + CC' + K_0 R K_0' \end{aligned}$$

We therefore have the distribution $x_1 \mid y^0 \sim N(\hat{x}_1, \Sigma_1)$.

We therefore have the recursive system

$$\begin{aligned} a_t &= y_t - G\hat{x}_t \\ K_t &= A_o \Sigma_t G' \left(G\Sigma_t G' + R\right)^{-1} \\ \hat{x}_{t+1} &= A_o \hat{x}_t + K_t a_t \\ \Sigma_{t+1} &= \left(A_o - K_t G\right) \Sigma_t \left(A_o - K_t G\right)' + CC' + K_t R K_t' \end{aligned} \tag{4.6}$$

- The first equation defines the **innovation** a_t, which is the deviation of the observed y_t from its best predictor $G\hat{x}_t$ constructed given y^{t-1}.
- The second equation defines the **Kalman gain**, which tells how much the innovation updates the previous best guess of the state $\hat{x}_{t+1}$
- The third equation is the law of motion for the mean forecast $\hat{x}_{t+1}$. Notice that the best forecast of $\hat{x}_{t+1}$ given y^{t-1} is $A_0\hat{x}_t$, to which we add $K_t a_t$ as the contribution of the information from y_t.
- Finally, we update the accuracy (variance) of the forecast Σ_{t+1}.

We can substitute for K_t into the law of motion for Σ_t to obtain

$$\Sigma_{t+1} = A_o \Sigma_t A_o' - A_o \Sigma_t G' \left(G \Sigma_t G' + R \right)^{-1} G \Sigma_t A_o' + CC'. \tag{4.7}$$

This is a matrix **Riccati equation** which often appears in linear-quadratic dynamic programming.

- The evolution of $\hat{x}_t$ is stochastic, being updated by the innovations that are constructed from the observations of y_t.
- The evolution of Σ_t deterministic. Σ_t typically converges to a constant in a time-invariant model, and the constant is zero when $CC' = 0$.
- This says that all observations y_t are equally informative, regardless of their particular value. This result is specific to this particular linear-Gaussian model.

The path $\hat{x}_t$ is often called the **filtered path** of x_t

- it represents the most likely location of x_t conditional on y^{t-1}

Can subsequent realizations of y_{t+j}, $j = 0, 1, 2, \ldots$ make the estimate of x_t more precise?

- They can. y_{t+j} is a signal about x_{t+j}, and knowledge where the state x is at time $t + j$ is also informative about where the state has been at time t.
- This is what the **Kalman smoother** does.

The linear-Gaussian model was used specifically to keep the filtering problem tractable.

Other such environments exist

- for example, filtering of the unknown state of the n-state Markov chain

Filtering methods are also applied in continuous-time environments

- Kalman and Bucy (1961) filter for filtering paths of Brownian motions
- Liptser and Shiryaev (2001) filters as a generalization of Kalman and Bucy (1961)
- Wonham (1964) filter for filtering regime shifts in continuous-time models

Additional informational frictions may be added.

- rational inattention/costly signal acquisition: Sims (2003)

Departures from Bayesian updating may be used to model behavioral features.

- adaptive expectations: Cagan (1956), Friedman (1957)
- least squares learning: Marcet and Sargent (1989)
- diagnostic expectations: Bordalo et al. (2020)
- extrapolation from past observations: Adam et al. (2016)
- memory loss: Azeredo da Silveira et al. (2022)

The literature is extensive, the key is how to discipline departures from Bayesian updating.

Applications

Muth's example

In the 1950's, Phillip Cagan (Cagan (1956)), Milton Friedman (Friedman (1957)), and others studied models of **adaptive expectations**

- expectations about the future slowly adjust in response to arrival of new data

In Muth (1960), John Muth asked what type of underlying stochastic processes would 'rationalize' the adaptive expectations model as the best statistical forecast of the future.

- this can be viewed a precursor of the assumption of rational expectations, more fully developed in Muth (1961)
- the solution to the problem is close to the filtering solution of Kalman (1960)
- the idea is to postulate a stochastic process under which the adaptive expectations model can be interpreted as the result of optimal learning (filtering)

Consider the model for agent's adaptive expectations

$$\begin{aligned} y^*_{t+1} &= K\sum_{j=0}^{\infty}(1-K)^j y_{t-j} \\ &= (1-K)\, y^*_t + K y_t \end{aligned} \tag{4.8}$$

- y_t is a time series we want to forecast, given its observations up to time t
- K is the weight on the current observation for the time-t forecast of y_{t+1}, denoted y^*_{t+1}.
- Cagan (1956): model of agent's forecasts of future inflation
- Friedman (1957) for forecasts of future income in a consumption-saving problem

Muth (1960) studied a model that can be written as a special case of the system in the Kalman filter problem:

$$\begin{aligned} x_{t+1} &= x_t + w_{t+1} \\ y_t &= x_t + v_t \end{aligned} \tag{4.9}$$

where w_t, v_t are independent scalar shocks with covariances Q and R, respectively, and y_t and x_t are also scalar. In the context of the Kalman filter model (4.1), we have

$$A_0 = 1,\ CC' = Q,\ G = 1$$

Then the filtering equations (4.6) together with (4.7) become

$$\begin{aligned} a_t &= y_t - \hat{x}_t \\ K_t &= \frac{\Sigma_t}{\Sigma_t + R} \\ \hat{x}_{t+1} &= \hat{x}_t + K_t a_t \\ \Sigma_{t+1} &= \Sigma_t - \frac{\Sigma_t^2}{\Sigma_t + R} + Q \end{aligned}$$

When we take the limit as $t \to \infty$, we expect $\Sigma_t \to \Sigma$ and $K_t \to K$.

Then the law of motion for the forecast is given by

$$\begin{aligned} \hat{x}_{t+1} &= \hat{x}_t + Ka_t = \hat{x}_t + K(y_t - \hat{x}_t) = (1-K)\hat{x}_t + Ky_t \\ &= \frac{R}{\Sigma + R}\hat{x}_t + \frac{\Sigma}{\Sigma + R}y_t \end{aligned}$$

- this forecasting formula is in line with the 'adaptive' forecast model (4.8)
- optimal filtering (best forecast) in the model (4.9) yields an belief updating formula which Cagan (1956) and Friedman (1957) interpreted as adaptive expectations.

We can map the motivating model of Jovanovic (1979) into our framework.

- θ is the unknown match quality, which can be treated as a constant state

$$\begin{aligned} \theta_{t+1} &= \theta_t = \theta \\ y_t &= \theta_t + v_t \end{aligned}$$

- the worker-firm pair has a prior $\theta \sim N(m_{-1}, \Sigma_0)$
- the time-t forecast of θ is denoted $m_t = \hat{x}_{t+1} = E\left[\theta \mid y^t\right]$
- the model then fits into the Kalman filter framework with $A_o = 1$, $C = 0$, $G = 1$, $R > 0$, and we thus obtain

$$\begin{aligned} a_t &= y_t - m_{t-1} \\ K_t &= \frac{\Sigma_t}{\Sigma_t + R} \\ m_t &= m_{t-1} + K_t a_t \\ \Sigma_{t+1} &= \frac{\Sigma_t R}{\Sigma_t + R}. \end{aligned}$$

This can be summarized as

$$\begin{aligned} m_t &= (1 - K_t)\, m_{t-1} + K_t y_t \\ K_t &= \frac{\Sigma_t}{\Sigma_t + R} \\ \frac{1}{\Sigma_{t+1}} &= \frac{1}{R} + \frac{1}{\Sigma_t} \end{aligned}$$

- the quantity Σ_t^{-1} is called precision of the forecast
- since $\Sigma_t^{-1} \to \infty$ over time, the value of the parameter θ is ultimately learned
- over time, the Kalman gain declines to zero, as additional observations become less and less informative
- this is contrary to the case when the unknown state x_t fluctuates over time

Summary

Summary

Information problems are essential in macroeconomics and finance.

- a key aspect is how to preserve tractability of the environment

The linear state space model preserves its tractability also under learning.

- the dynamics under learning continue to be linear
- this means that the model can easily be embedded in models solved by perturbation approximations etc.

The hidden state space model is heavily featured in estimation.

- many macroeconomic models have a Markov solution with a Markov state that is not observable
- the econometrician observes macroeconomic data that are an imperfect reflection of the underlying state
- part of the estimation involves filtering the best estimation of the path of the Markov state

Appendix

Adam, Klaus, Albert Marcet, and Juan Pablo Nicolini (2016) "Stock Market Volatility and Learning," *Journal of Finance*, 71 (1), 33–82.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer (2020) "Overreaction in Macroeconomic Expectations," *American Economic Review*, 110 (9), 2748–2782.

Cagan, Phillip D. (1956) "The Monetary Dynamics of Hyperinflation," in Friedman, Milton ed. *Studies in the Quantity Theory of Money*, 25–117: University of Chicago Press.

Friedman, Milton (1957) *A Theory of the Consumption Function*: Princeton University Press, Princeton.

Jovanovic, Boyan (1979) "Job Matching and the Theory of Turnover," *Journal of Political Economy*, 87 (5), 972–990.

Kalman, Rudolph Emil (1960) "New Approach to Linear Filtering and Prediction Problems," *Transactions of the ASME–Journal of Basic Engineering*, 82 (Series D), 35–45.

Kalman, Rudolph Emil and Richard Snowden Bucy (1961) "New Results in Linear Filtering and Prediction Theory," *Journal of Basic Engineering*, 83 (1), 95–108.

Liptser, Robert S. and Albert N. Shiryaev (2001) *Statistics of Random Processes*: Springer Verlag, New York, 2nd edition.

Ljungqvist, Lars and Thomas J. Sargent (2020) "Recursive Macroeconomic Theory," Unpublished manuscript, draft of 5th edition.

Marcet, Albert and Thomas J. Sargent (1989) "Convergence of Least Squares Learning Mechanisms in Self-Referential Linear Stochastic Models," *Journal of Economic Theory*, 48 (2), 337–368.

Muth, John F. (1960) "Optimal Properties of Exponentially Weighted Forecasts," *Journal of the American Statistical Association*, 55 (290), 299–306.

——— (1961) "Rational Expectations and the Theory of Price Movements," *Econometrica*, 29 (3), 315–335.

Azeredo da Silveira, Rava, Yeji Sung, and Michael Woodford (2022) "Optimally Imprecise Memory and Biased Forecasts," Working paper, Columbia University.

Sims, Christopher A. (2003) "Implications of Rational Inattention," *Journal of Monetary Economics*, 50 (3), 665–690.

Wonham, W. Murray (1964) "Some Applications of Stochastic Differential Equations to Optimal Nonlinear Filtering," *SIAM Journal on Control and Optimization*, 2 (3), 347–369, 10.1137/0302028.