Chapter 9

Learning

9.1 Introduction

We begin with a Markov process, but we suppose that either parameters are unknown or more generally that the states are not observed. We then construct another Markov process that summarizes information about the hidden state that is contained in the history of the signals. The state vector in this new process consists of a set of sufficient statistics for the probability distribution of the original hidden Markov state conditional on the history of signals.

We study recursive learning algorithms that implement Bayes' rule for four versions of a Hidden Markov Model. These feature learning about

1. Fixed parameters (invariant states)
2. Discrete states
3. Multiple VAR regimes
4. A continuously distributed state vector in a linear system

9.2 Learning Parameters

Our first problem is to learn unknown parameters of a linear regression model that are regarded as constant hidden states. A Bayesian statistician assigns a prior probability distribution to statistical models indexed by parameter values. Probabilistically mixing statistical models in this way leads to a stochastic process described by Proposition 2.7.1 that is not ergodic. We want to make probability

assessments that do not condition on invariant events. In this setting, conditioning on invariant events means knowing parameters. We use the conjugate prior analysis of Luce and Raiffa (1957) to learn about parameters sequentially from increments to a data history.

Consider the first-order vector autoregressive model:

$$\begin{aligned} X_{t+1} &= AX_t + BW_{t+1} \\ Y_{t+1} - Y_t &= H + DX_t + FW_{t+1} \end{aligned} \tag{9.1}$$

with an observable state vector but unknown parameters. Suppose that $Y_{t+1} - Y_t$ and W_{t+1} have the same dimensions and that F is nonsingular. Also suppose that X_t consists of $Y_t - Y_{t-1} - H$ and a finite number of lags of this vector.

Insights of Zellner (1962), Box and Tiao (1992), Sims and Zha (1999), and especially Zha (1999) allow us transform system (9.1) in a way that facilitates making valid statistical inferences by applying least squares equation by equation. Factor the matrix $FF' = J\Delta J'$, where J is lower triangular with ones on the diagonal and Δ is diagonal. Construct

$$J^{-1}(Y_{t+1} - Y_t) = J^{-1}H + J^{-1}DX_t + U_{t+1} \tag{9.2}$$

where

$$U_{t+1} = J^{-1}FW_{t+1}$$

and hence has Δ as its covariance matrix. Moreover, the i^{th} entry of U_{t+1} is uncorrelated with and consequently independent of the jth entries of $Y_{t+1} - Y_t$ for $j = 1, 2, ..., i-1$. As a consequence, each equation in system (9.2) can be interpreted as a regression equation in which the left-hand side variable in equation i is the i^{th} entry of $Y_{t+1} - Y_t$. The regressors are a constant, the entries of X_t, and the j^{th} entries of $Y_{t+1} - Y_t$ for $j = 1, \ldots, i-1$. We think of each of these equations as unrestricted regression equations with disturbance terms that are uncorrelated across the equations. This system is recursive. The first equation determines the first entry of $Y_{t+1} - Y_t$, the second equation determines the second entry of $Y_{t+1} - Y_t$ given the first entry, and so forth. We can build the coefficients A, B, D, F, H from the constructed regression equations with one caveat. Knowledge of J and Δ determine FF' only for a nonsingular F. One solution is $F = J\Delta^{1/2}$, where a diagonal matrix raised to the one-half power is built by taking the square root of each diagonal entry. Because other solutions F exist, F is not identified without additional restrictions.

Consider, in particular, the ith such regression formed in this way, which we express as the scalar regression model:

$$Y_{t+1}^{[i]} - Y_t^{[i]} = {R_{t+1}^{[i]}}' \beta^{[i]} + U_{t+1}^{[i]}$$

where $R_{t+1}^{[i]}$ is the appropriate vector of regressors in the ith equation of system (9.2). To simplify notation, we will omit the superscripts and understant that we are estimating one equation at a time. The disturbance U_{t+1} is a normally distributed random vector with mean zero and variance σ^2. Moreover, U_{t+1} is independent of R_{t+1}. Suppose that $Y_{t+1} - Y_t$ is observed along with R_{t+1} as of date $t+1$ but β or σ^2 are unknown. We let $Y^t = [(Y_t - Y_{t-1})', \ldots, (Y_1 - Y_0)']'$ and X_0 denote the information that is observed as of date t.

Let the distribution of β conditioned on Y^t, X_0, and σ^2 be normal with mean b_t and precision matrix $\zeta\Lambda_t$ where $\zeta = \frac{1}{\sigma}^2$ and the precision matrix is the inverse of a conditional covariance matrix of the unknown parameters. Including the new date $t+1$ information $Y_{t+1} - Y_t$, it follows that the distribution of β conditioned on Y^{t+1}, X_0, and σ^2 is also normal but with precision $\zeta\Lambda_{t+1}$, where

$$\Lambda_{t+1} = R_{t+1}R_{t+1}{}' + \Lambda_t \tag{9.3}$$

$\zeta = \frac{1}{\sigma^2}$. Since $\Lambda_{t+1} - \Lambda_t$ is a positive semidefinite matrix, this updating equation is consistent with the notion that additional information improves accuracy. Thus, Λ_{t+1} cumulates cross-products of the regressors and adds them to an initial Λ_0. The conditional mean update for the normal distribution of unknown coefficients can be deduced from Λ_{t+1} via the updating equation:

$$\Lambda_{t+1}b_{t+1} = \left[\Lambda_t b_t + R_{t+1}(Y_{t+1} - Y_t)\right]. \tag{9.4}$$

Notice how $\Lambda_{t+1}b_{t+1}$ cumulates cross-products of regressors and the left side variable of the regression, then adds the outcome to an initial condition.

So far we have conditioned on σ^2, which is equivalent to conditioning on its inverse ζ. Assume a date t gamma density for ζ conditioned on Y^t, X_0:

$$\propto (\zeta)^{\frac{c_t}{2}} \exp(-d_t\zeta/2),$$

where the density is expressed as a function of ζ, so that $d_t\zeta$ has a chi-square density with $c_t + 1$ degrees of freedom. The implied density for ζ conditioned on time $t+1$ information has the same functional form with updated parameters:

$$\begin{aligned} c_{t+1} &= c_t + 1 \\ d_{t+1} &= (Y_{t+1} - Y_t)^2 - (b_{t+1})'\Lambda_{t+1}b_{t+1} + (b_t)'\Lambda_t b_t + d_t. \end{aligned}$$

The distribution of β conditioned on Y^{t+1}, X_0, and ζ is normal with mean b_{t+1} and precision matrix $\zeta\Lambda_{t+1}$; the distribution for ζ conditioned on Y^{t+1}, X_0 has a gamma density:

$$\propto (\zeta)^{\frac{c_{t+1}}{2}} \exp(-d_{t+1}\zeta/2).$$

A decision-maker who does not know the underlying parameters continues to have a Markov decision problem except that b_t, c_t, d_t must now be included along with the state vector X_t.

One special choice of prior probability distributions implies standard least squares regression statistics. In particular, to express priors that are not informative, it is common to use so called "improper priors" that do not integrate to unity.[1] First set $\Lambda_0 = 0$, which in effect imposes a uniform but improper prior over β. The implied normal distributions will be proper after we have accumulated enough observations to make Λ_{t+1} become nonsingular. Although Λ_t's early in the sequence are singular, nevertheless we can update $\Lambda_{t+1}b_{t+1}$ via (9.4); b_{t+1} will not be uniquely defined until Λ_{t+1} becomes nonsingular. When $\Lambda_0 = 0$ the specification of b_0 is inconsequential, and b_{t+1} becomes the standard least squares estimator. The "improper gamma" prior often associated with our previous improper normal prior sets c_0 to minus two and d_0 to zero. This is obtained by imposing a uniform prior for the logarithm of the precision ζ or for the logarithm of σ^2. With this combination of priors, d_{t+1} is the sum of squared regression residuals.[2]

Recall that in figure 42 we showed that the long-term consumption response to a permanent shock was about double that of the short-term response. **?** used the method just described to assess the accuracy of these measurements. The short-term impact of shocks is measured with much more accuracy, as Figure 91 makes evident. We take this evidence to be that there *could be* a long-term risk component to consumption, but that it is poorly measured.

9.3 Learning Discrete States

Suppose that $\{X_t\}$ evolves as an n-state Markov process with transition matrix $\mathbb{P}$. There is a vector of signals denoted by $Y_{t+1} - Y_t$ with density $\psi_i(y^*)$ if state i is realized, meaning that X_t is the i^{th} coordinate vector. We want to compute the probability of being in state i given the signal history. The vector of probabilities is $Q_t = E\left[X_t | Y^t, Q_0\right]$ where Q_0 is the vector of initial probabilities. We proceed recursively through the following steps:

i) Find the joint distribution for $(X_{t+1}, Y_{t+1} - Y_t)$ conditioned on X_t. The conditional distributions for $Y_{t+1} - Y_t$ and X_{t+1} are independent by assumption.

[1] Such a procedure can result in estimators that are inadmissible.

[2] See Box and Tiao (1992) for a discussion of improper priors including the specification given here for the regression model.

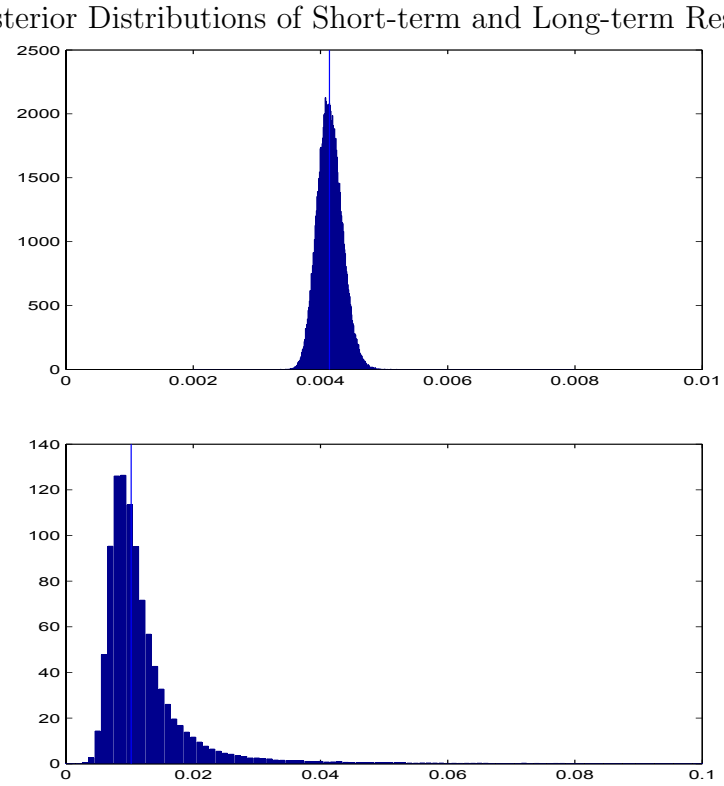

Figure 91: This figure plots posterior histograms for the magnitudes of the short-term and long-term responses of the logarithm of consumption to the shocks. The magnitude is measured as the absolute value across the contributions from the two shocks. The upper panel depicts the histogram for the immediate response and the lower panel depicts the histogram for the long-term limiting response. The figure comes from Hansen et al. (2008).

Write the joint density conditioned on X_t as:

$$\begin{array}{ccc} (\mathbb{P}'X_t) & \times & (X_t)'\text{vec}\left\{\psi_i(y^*)\right\} \\ \uparrow & & \uparrow \\ X_{t+1} \text{ density} & & Y_{t+1} - Y_t \text{ density} \end{array} \tag{9.5}$$

where $\text{vec}(r_i)$ is a column vector with r_i in the i^{th} entry. We have imposed conditional independence by forming a joint conditional distribution as a product of two conditional densities, one for X_{t+1} and one for $Y_{t+1} - Y_t$.

ii) Find the joint distribution of $X_{t+1}, Y_{t+1} - Y_t$ conditioned on Q_t. Since X_t is not observed, we form the appropriate average of (9.5) conditioned on Y^t, Q_0:

$$\mathbb{P}' \text{diag}\{Q_t\} \text{ vec}\{\psi_i(y^*)\}, \tag{9.6}$$

where $\text{diag}(Q_t)$ is a diagonal matrix with the entries of Q_t on the diagonal. Thus, Q_t encodes all information in the history of signals that is pertinent for this calculation. Notice that conditioned on Q_t, the distributions for X_{t+1} and $Y_{t+1} - Y_t$ are *not* independent.

iii) Find the implied distribution for $Y_{t+1} - Y_t$ conditioned on Q_t. Summing (9.6) over the hidden states gives

$$(\mathbf{1}_n)' \mathbb{P}' \text{ diag}\{Q_t\} \text{vec}\{\psi_i(y^*)\} = Q_t \cdot \text{vec}\{\psi_i(y^*)\}.$$

Thus, Q_t is a vector of weights used in forming a mixture distribution. Suppose, for instance, that ψ_i is a normal distribution with mean with mean μ_i and covariance matrix Σ_i. Then the distribution of $Y_{t+1} - Y_t$ conditioned on Q_t is a *mixture of normals* with mixing probabilities given by appropriate entries of Q_t.

iv) Obtain Q_{t+1} by dividing the *joint* density for $(Y_{t+1} - Y_t, X_{t+1})$ conditioned on Q_t by the *marginal* density for $Y_{t+1} - Y_t$ conditioned on Q_t and then evaluating this ratio at $Y_{t+1} - Y_t$. Division gives the density for X_{t+1} conditioned $(Q_t, Y_{t+1} - Y_t)$, which in this case is just a vector Q_{t+1} of conditional probabilities. Thus, we are led to

$$Q_{t+1} = \left(\frac{1}{Q_t \cdot \text{vec}\{\psi_i(Y_{t+1} - Y_t)\}} \right) \mathbb{P}' \text{diag}(Q_t) \text{vec}\{\psi_i(Y_{t+1} - Y_t)]\} \tag{9.7}$$

Taken together, steps (iii) and (iv) define a Markov process for Q_{t+1}. As indicated in step iii, $Y_{t+1} - Y_t$ is is drawn from a (history dependent) mixture of densities ψ_i; and as indicated in step (iv), the vector Q_{t+1} is the exact function of $Y_{t+1} - Y_t$, Q_t given in (9.7).

9.4 Multiple VAR Regimes

Consider again Example 8.2.4 except, following Sclove (1983) and Hamilton (1989), suppose that there are multiple VAR regimes (A_i, B_i, D_i, F_i) (with F_i nonsingular) for $i = 1, 2, ..., n$, where the indices i are governed by a Markov process with transition matrix $\mathbb{P}$. We can think of X_t and a regime indicator W_t

jointly as a Markov process. The realized values of W_t is a coordinate vector with a one in the i^{th} coordinate if regime i is realized. Consider the case in which W_t is not observed. Let Q_t denote an n-dimensional vector of probabilities over the hidden states conditioned on Y^t, X_0, and Q_0, where Q_0 is the date zero vector of initial probabilities for W_0. Equivalently, Q_t is $E(W_t|Y^t, X_0, Q_0)$ where Q_0 is the date zero vector of initial probabilities for W_0. The vector Q_t solves a *filtering problem*. We can represent (X_t, Q_t) as a Markov process by executing the following three steps.

i) Find the joint distribution for $(W_{t+1}, Y_{t+1} - Y_t)$ conditioned on (W_t, X_t). The conditional distributions for W_{t+1} and $Y_{t+1} - Y_t$ are independent by assumption. Conditioned on W_t, X_t conveys no information about W_{t+1} and thus the conditional density for W_{t+1} is given by entries of $\mathbb{P}'W_t$. Conditioned on $W_t = i$, $Y_{t+1} - Y_t$ is normal with mean D_iX_t and covariance matrix $F_i(F_i)'$. Let $\psi_i(y^*, X_t)$ be the corresponding normal density function for $Y_{t+1} - Y_t$ conditioned on X_t when W_t is in regime i. We write the joint density conditioned on (X_t, W_t) as:

$$\underbrace{\left(\mathbb{P}'W_t\right)}_{\substack{\uparrow \\ W_{t+1} \text{ density}}} \quad \times \quad \underbrace{(W_t)'\text{vec}\left\{\psi_i(y^*, X_t)\right\}}_{\substack{\uparrow \\ Y_{t+1} - Y_t \text{ density}}} \tag{9.8}$$

where $\text{vec}(r_i)$ is a column vector with r_i in the i^{th} entry. We have imposed conditional independence by forming joint conditional distribution as a product of two conditional densities, one for W_{t+1} and one for $Y_{t+1} - Y_t$.

ii) Find the joint distribution for $W_{t+1}, Y_{t+1} - Y_t$ conditioned on (X_t, Q_t). Since W_t is not observed, we form the appropriate average of (9.8) conditioned on the Y^t, X_0, Q_0:

$$\mathbb{P}'\text{diag}\{Q_t\}\ \text{vec}\left\{\psi_i(y^*, X_t)\right\} \tag{9.9}$$

where $\text{diag}(Q_t)$ is a diagonal matrix with the entries of Q_t on the diagonal. Thus, Q_t encodes the information in Y^t, X_0 and Q_0 that is pertinent for this calculation. Notice that conditioned on (X_t, Q_t), the distributions for $Y_{t+1} - Y_t$ and W_{t+1} are *not* independent.

iii) Find the distribution of $Y_{t+1} - Y_t$ conditioned on (X_t, Q_t). Summing (9.9) over the hidden states gives

$$(\mathbf{1}_n)'\mathbb{P}'\ \text{diag}\{Q_t\}\text{vec}\left\{\psi_i(y^*, X_t)]\right\} = Q_t \cdot \text{vec}\left\{\psi_i(y^*, X_t)\right\},$$

Thus, the distribution for $Y_{t+1} - Y_t$ conditioned on (X_t, Q_t) is a *mixture of normals* in which the probability that $Y_{t+1} - Y_t$ is normal with mean D_iX_t

and covariance matrix $F_i F_i'$ with probability given by the i^{th} entry of Q_t. Similarly, the distribution of X_{t+1} is a mixture of normals.

iv) Obtain Q_{t+1} by dividing the *joint* density for $(Y_{t+1} - Y_t, W_{t+1})$ conditioned on (X_t, Q_t) by the *marginal* density for $Y_{t+1} - Y_t$ conditioned on (X_t, Q_t). Division gives the density for W_{t+1} conditioned $(Y_{t+1} - Y_t, X_t, Q_t)$, which in this case is just a vector Q_{t+1} of conditional probabilities. Thus, we are led to

$$Q_{t+1} = \left(\frac{1}{Q_t \cdot \text{vec}\left\{\psi_i(Y_{t+1} - Y_t, X_t)\right\}} \right) \mathbb{P}' \text{diag}(Q_t) \text{vec}\left\{\psi_i(Y_{t+1} - Y_t, X_t)\right\} \tag{9.10}$$

Taken together, steps (iii) and (iv) provide a Markov evolution for (X_{t+1}, Q_{t+1}). As argued in (iii), $Y_{t+1} - Y_t$ is a mixture of normally distributed random variables; and as argued in step three, the vector Q_{t+1} is an exact function of $Y_{t+1} - Y_t$, Q_t, and X_t that is given by formula (9.10).

9.5 Linear Filtering

We now consider another extension of Example 8.2.4 by assuming that there is only one VAR regime and that the state X_t is not fully observed. We construct a multiplicative likelihood process in a setting in which part of the state is hidden. Thus, the state and observations are related by a linear state space system

$$\begin{aligned} X_{t+1} &= AX_t + BW_{t+1} \\ Y_{t+1} - Y_t &= H + DX_t + FW_{t+1}, \end{aligned}$$

where W_{t+1} is a standard normally distributed random vector independent of date t information including the current hidden state, X_t is a (partially) hidden state, $Y_{t+1} - Y_t$ is observed at $t+1$, and $X_0 \sim \mathcal{N}(\bar{X}_0, \Sigma_0)$. We let Q_0 denote this initial distribution. We do not restrict A to be stable and allow for there to be an invariant state to be learned over time.

While $\{X_t, t = 0, 1, 2, \ldots\}$ is Markov, $\{Y_{t+1} - Y_t, t = 0, 1, 2, \ldots\}$ is not. We form a new Markov process in which the date t state is Q_t, the distribution of the current Markov state X_t given Y^t and Q_0. The state vector Q_t is a distribution function indexed by t. For this example, the distribution Q_t is multivariate normal, so it suffices to keep track only of the mean $\bar{X}_t$ and covariance matrix Σ_t of X_t conditioned on Q_0 and Y^t. Thus, $\bar{X}_t$ and Σ_t are sufficient statistics for the history Y^t and Q_0. The Kalman filter constructs Q_t recursively.

While the distribution Q_t is interesting in its own right, it is also a key input into the likelihood process

$$L_{t+1} = \prod_{j=1}^{t+1} \int \psi(Y_j - Y_{j-1}|x) Q_{j-1}(dx) \tag{9.11}$$

given the matrices A, B, D, F.

We compute the Kalman filter in the following three step process.

i) Note that the joint distribution of $X_{t+1}, Y_{t+1} - Y_t$ conditional on X_t is

$$\begin{bmatrix} X_{t+1} \\ Y_{t+1} - Y_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} A \\ D \end{bmatrix} X_t, \begin{bmatrix} B \\ F \end{bmatrix} \begin{bmatrix} B' & F' \end{bmatrix} \right).$$

ii) Let Q_t be the normal distribution function with mean $\overline{X}_t$ and covariance matrix Σ_t; Q_t is the distribution of X_t conditioned on Y^t and Q_0. Note that because

$$\begin{bmatrix} X_{t+1} \\ Y_{t+1} - Y_t \end{bmatrix} = \begin{bmatrix} 0 \\ H \end{bmatrix} + \begin{bmatrix} A \\ D \end{bmatrix} \overline{X}_t + \begin{bmatrix} A \\ D \end{bmatrix} (X_t - \overline{X}_t) + \begin{bmatrix} B \\ F \end{bmatrix} W_{t+1},$$

it follows that the joint distribution of $X_{t+1}, Y_{t+1} - Y_t$ conditioned on Y^t and Q_0 is

$$\begin{bmatrix} X_{t+1} \\ Y_{t+1} - Y_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ H \end{bmatrix} + \begin{bmatrix} A \\ D \end{bmatrix} \bar{X}_t, \begin{bmatrix} A \\ D \end{bmatrix} \Sigma_t \begin{bmatrix} A' & D' \end{bmatrix}' + \begin{bmatrix} B \\ F \end{bmatrix} \begin{bmatrix} B' & F' \end{bmatrix} \right)$$

and in particular that the marginal distribution of $Y_{t+1} - Y_t$ conditional on Q_t is

$$Y_{t+1} - Y_t \sim \mathcal{N}(H + D\bar{X}_t, D\Sigma_t D' + FF').$$

This step gives $\phi(y^*|Q_t)$.

iii) Compute the density of X_{t+1} conditional on $Y_{t+1} - Y_t$ and Q_t by dividing the joint distribution for $(X_{t+1}, Y_{t+1} - Y_t)$ conditioned on Q_t by the marginal density for $Y_{t+1} - Y_t$ conditional on Q_t. The Gaussian structure implies that we can compute this conditional distribution by running a population linear least squares regression of $X_{t+1} - A\overline{X}_t$ on $Y_{t+1} - Y_t - D\overline{X}_t$, namely,

$$E[(X_{t+1} - A\overline{X}_t)|Y_{t+1} - Y_t - H - D\overline{X}_t, Q_t] = \mathcal{K}(\Sigma_t)(Y_{t+1} - H - D\overline{X}_t)$$

where

$$\mathcal{K}(\Sigma_t) = (A\Sigma_t D' + BF')(D\Sigma_t D' + FF')^{-1}. \tag{9.12}$$

Then Q_{t+1} is constructed as

$$X_{t+1}|Y_{t+1} - Y_t, Q_t \sim \mathcal{N}(\overline{X}_{t+1}, \Sigma_{t+1})$$

where

$$\overline{X}_{t+1} = A\overline{X}_t + \mathcal{K}(\Sigma_t)(Y_{t+1} - Y_t - H - D\overline{X}_t)$$

and

$$\begin{aligned}\Sigma_{t+1} =& A\Sigma_t A' + BB' \\ & -(A\Sigma_t D' + BF')(D\Sigma_t D' + FF')^{-1}(D'\Sigma_t A + F'B). \end{aligned} \tag{9.13}$$

This gives Q_{t+1} as an exact function of $Y_{t+1} - Y_t$ and Q_t.

Thus, steps ii and iii taken together give the evolution of $\{Y_{t+1} - Y_t, Q_{t+1})\}$ as a first-order Markov process. These caluculations together with the initial distribution Q_0 for $X_0 \sim \mathcal{N}(\overline{X}_0, \Sigma_0)$ give a recursive representation (9.11) for the likelihood process for $\{Y_t, t = 1, 2, \ldots\}$. Equations (9.12) and (9.13) are the celebrated Kalman filtering equations.

If Σ_0 is a fixed point of iterations on (9.12), (9.13), $\mathcal{K}(\Sigma_t) = K$ for all $t \geq 1$, further simplifying the recursive representation (9.11). Setting Σ_0 to the positive semidefinite fixed point amounts to pretending that initially we are conditioning on an infinite history of Y's.