T. Rothenberg Fall, 2007

State Space Models and the Kalman Filter

1 Introduction

Many time-series models used in econometrics are special cases of the class of linear state space models developed by engineers to describe physical systems. The Kalman filter, an efficient recursive method for computing optimal linear forecasts in such models, can be exploited to compute the exact Gaussian likelihood function.

The linear state-space model postulates that an observed time series is a linear function of a (generally unobserved) state vector and the law of motion for the state vector is first-order vector autoregression. More precisely, let y_t be the observed variable at time t and let α_t denote the values taken at time t by a vector of p state variables. Let A and b be $p \times p$ and $p \times 1$ matrices of constants. We assume that $\{y_t\}$ is generated by

$$\begin{aligned} y_t &= b'\alpha_t + u_t, && (1) \\ \alpha_t &= A\alpha_{t-1} + v_t && (2) \end{aligned}$$

where the scalar u_t and the vector v_t are mean zero, white-noise processes, independent of each other and of the initial value α_0. We denote $\sigma^2 = E(u_t^2)$ and $\Sigma = E(v_t v_t')$. Equation (1) is sometimes called the "measurement" equation while (2) is called the "transition" equation. The assumption that the autoregression is first-order is not restrictive, since higher-order systems can be handled by adding additional state variables.

In most engineering (and some economic) applications, the α's represent meaningful but imperfectly measured physical variables. Models based on the "permanent" income hypothesis are classic examples. But sometimes state-space models are used simply to exploit the fact that rather complicated dynamics in an observable variable can result from adding noise to a linear combination of autoregressive variables. For example, all ARMA models for y_t can be put in state space form even though the state variables α_t have no particular economic meaning. An even richer class of (possibly nonstationary) state space models can be produced by introducing an observed exogenous forcing variable x_t into the measurement equation, by letting b, A, σ^2, and Σ depend on t, and by letting y_t be a vector. Since these generalizations complicate the notation but do not affect the basic theory, they will be ignored in these notes.

2 ARMA Models in State Space Form

Consider the ARMA(1,1) model

$$y_t = \varphi y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}.$$

Defining $\alpha_t = (\alpha_{1t}, \alpha_{2t})' = (y_t, \theta\varepsilon_t)'$, we can write $y_t = b'\alpha_t$ where $b = (1, 0)'$ and

$$\begin{bmatrix} \alpha_{1t} \\ \alpha_{2t} \end{bmatrix} = \begin{bmatrix} \varphi & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_{1,t-1} \\ \alpha_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \theta\varepsilon_t \end{bmatrix}.$$

Thus the ARMA(1,1) model has a state-space representation with $u_t = 0$.

More generally, suppose $\{y_t\}$ is a mean-zero ARMA(p,q) process. Let $m = \max(p, q+1)$. Then, we can write

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_m y_{t-m} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_{m-1}\varepsilon_{t-m+1}$$

with the redundant coefficients set to zero. Define the column vectors

$$\underset{m\times 1}{b} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \underset{(m-1)\times 1}{c} = \begin{bmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{m-1} \end{bmatrix}, \qquad \underset{m\times 1}{d} = \begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_{m-1} \end{bmatrix}.$$

By successive substitution, one can verify that y_t has the state space representation

$$y_t = b'\alpha_t, \qquad \alpha_t = A\alpha_{t-1} + v_t$$

where α_t is an m-dimensional state vector, $u_t = 0, v_t = d\varepsilon_t$ and

$$A = \begin{bmatrix} c & I_{m-1} \\ \varphi_m & \mathbf{0}' \end{bmatrix}.$$

3 The Kalman Filter

Denote the vector $(y_1, ..., y_t)$ by Y_t .The Kalman filter is a recursive algorithm for producing optimal linear forecasts of α_{t+1} and y_{t+1} from the past history Y_t, assuming that A, b, σ^2, and Σ are known. Define

$$a_t = E(\alpha_t|Y_{t-1}) \quad \text{and} \quad V_t = var(\alpha_t|Y_{t-1}). \tag{3}$$

If the u's and v's are normally distributed, the minimum MSE forecast of y_{t+1} at time t is $b'a_{t+1}$. The key fact (which we shall derive below) is that, under normality, a_{t+1} can be calculated recursively by

$$a_{t+1} = Aa_t + AV_tb\frac{y_t - b'a_t}{b'V_tb + \sigma^2}, \qquad V_{t+1} = \Sigma + AV_tA' - \frac{AV_tbb'V_tA'}{b'V_tb + \sigma^2} \tag{4}$$

starting with the appropriate initial values (a_1, V_1). To forecast $y_{t+1} = b'a_{t+1}$ at time t, one needs only the current y_t and the previous forecast of α_t and its variance. Previous values $y_1, ..., y_{t-1}$ enter only through a_t . Note that y_t enters linearly into the calculation of a_t and does not enter at all into the calculation of V_t . The forecast of y_t is a linear filter of previous y's. If the errors are not normal, the forecasts produced from iterating (4) are still of interest; they are best linear predictors.

The appropriate starting values a_1 and V_1 depend on the assumption made on α_0. If the $\{\alpha_t\}$ are covariance stationary, then each α_t must have zero mean and constant variance. In that case, $a_1 = E[\alpha_1] = 0$ and $V_1 = var[\alpha_1]$ must satisfy $V_1 = AV_1A' + \Sigma$. This implies

$$vec(V_1) = [I - (A \otimes A)]^{-1}vec(\Sigma). \tag{5}$$

In practice, one often uses mathematically convenient initial conditions and relies on the fact that, for weakly dependent processes, initial conditions do not matter very much. For more details, see A. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter* (1989), Chapter 3.

4 Using the Kalman Filter to Compute ML Estimates

Suppose we wish to estimate the unknown parameters of a given state-space model from the observations $y_1, ..., y_T$. Let $f(y_t|Y_{t-1})$ represent the conditional density of y_t, given the previous y's. The joint density function for the y's can always be factored as

$$f(y_1)f(y_2|Y_1)f(y_3|Y_2)...f(y_T|Y_{T-1}).$$

If the y's are normal, it follows from equations (1) and (2) that $f(y_t|Y_{t-1})$ is also normal with mean $b'a_t$ and variance $\sigma^2 + b'V_tb$. Hence, the log likelihood function is (apart from a constant)

$$-\frac{1}{2}\sum_{t=1}^{T}[\ln(b'V_tb+\sigma^2)+\frac{(y_t-b'a_t)^2}{b'V_tb+\sigma^2}] \tag{6}$$

and can be computed from the output of the Kalman filter. Of course, an alternative expression for the normal log-likelihood is

$$-\frac{1}{2}[\ln|\Omega|+y'\Omega^{-1}y]$$

where $y = (y_1, ..., y_T)'$ and $\Omega = E(yy')$. Thus, the Kalman filter can be viewed as a recursive algorithm for computing Ω^{-1} and $|\Omega|$. After evaluating the normal likelihood (for any given values of the parameters), quasi maximum likelihood estimates can be obtained by grid search or iterative methods such as employed in the Newton-Raphson algorithm.

The Kalman filter can also be used to compute GLS regression estimates. As an example, consider the regression model $y_t = \beta' x_t + u_t$, where x_t is a vector of K exogenous variables and u_t is a stationary normal ARMA(p,q) process with *known* parameters. Direct use of GLS requires finding the inverse of the variance matrix for the $u's$. This can be achieved more easily using the Kalman filter. If u_t were observable, one could put the model for u_t in state space form and compute via the Kalman filter the best linear predictor of u_t given its past history, say E(u_t|past u's) $= b'a_t$, and the prediction error variance Var(u_t|past u's) $= b'V_tb + \sigma^2$. Note that the T random variables

$$u_t^* = \frac{u_t - E(u_t|\text{past u'}s)}{\sqrt{Var(u_t|\text{past } u\text{'s})}} \qquad\qquad t = 1, ..., T$$

are uncorrelated with unit variance. Since $E(u_t|$past u's) is linear in past u's and $Var(u_t|$past u's) does not depend on the u_t at all, we can write in vector notation $u^* = Ru$ where R is a nonrandom triangular matrix. Of course, we do not observe the u's. But we can apply this filter to the y and X data, constructing $K + 1$ new time series $y^* = Ry$ and $X^* = RX$. Note that $y^* = X^*\beta + u^*$. If we regress y^* on X^*, the resulting coefficient is the GLS estimate since by construction u^* is white noise.

5 Derivation of the Recursion Equations

Recall that, if a scalar random variable Z and a random vector X are jointly normal, then

$$E(X|Z) = E(X) + \frac{cov(X,Z)}{var(Z)}(Z - EZ), \quad Var(X|Z) = Var(X) - \frac{cov(X,Z)cov(X,Z)'}{var(Z)} \tag{7}$$

Define the random variables $a_t^* = E(\alpha_t|Y_t)$ and $V_t^* = var(\alpha_t|Y_t)$. Note that a_t^* and a_t are both expectations of the same random variable α_t, the former conditioning on y_t and the latter not. Likewise V_t^* and V_t are both variances of α_t, the former conditioning on y_t and the latter not. Since, conditional on Y_{t-1}, the vector α_t and the scalar y_t are jointly normal, we can use (7) to calculate a relationship between a_t^* and a_t and between V_t^* and V_t. From (1) and (2) we have

$$\begin{aligned} Cov(\alpha_t, y_t|Y_{t-1}) &= Cov(\alpha_t, \alpha_t' b|Y_{t-1}) = V_tb \\ Var(y_t|Y_{t-1}) &= Var(b'\alpha_t + u_t|Y_{t-1}) = b'V_tb + \sigma^2 \\ E(\alpha_t|Y_{t-1}) &= a_t; \qquad E(y_t|Y_{t-1}) = b'a_t \end{aligned}$$

Thus, letting α_t play the role of X and y_t the role of Z, we have from (7)

$$a_t^* = a_t + V_t b \frac{y_t - b' a_t}{b' V_t b + \sigma^2} \quad \text{and} \quad V_t^* = V_t - \frac{V_t b b' V_t'}{b' V_t b + \sigma^2} \tag{8}$$

From (2), it follows that

$$a_{t+1} = A a_t^* \quad \text{and} \quad V_{t+1} = A V_t^* A' + \Sigma \tag{9}$$

The "updating" equations (8) describe how the forecast of the state vector at time t is changed when y_t is observed. Together with the "prediction" equations (9), they imply the recursion (4).

In models where the state variables have an economic interpretation, it is sometimes desirable to estimate α_t using all the available data. Starting with a_T and V_T computed with the Kalman filter, one can iterate backwards to compute $E(\alpha_t|Y_T)$. The relevant recursion, called the "smoothing" algorithm, is derived and discussed in Harvey's book.

6 Matrices that Diagonalize the Covariance Matrix for y

Again, let Y_t denote the vector $y_1, ..., y_t$. Note that a_t is a linear function of the data in Y_{t-1} and hence the prediction error $e_t = y_t - b' a_t$ is a linear function of the data in Y_t . If $t > s$,

$$E(e_t e_s) = E e_s E(b' \alpha_t - b' a_t + u_t | Y_{t-1}) = 0.$$

If $t = s$,

$$E e_t^2 = E[var(b' \alpha_t + u_t | Y_{t-1})] = \sigma^2 + b' V_t b.$$

Thus, the $\{e_t\}$ are a set of uncorrelated, but heteroskedastic random variables. Denoting the vector of the y's by $\mathbf{y}$ and the vector of the e's by $\mathbf{e}$, we have $\mathbf{e} = \mathbf{G}\mathbf{y}$, where $\mathbf{G}$ is a nonrandom triangular matrix such that $E\mathbf{e}\mathbf{e}' = \mathbf{G}(E\mathbf{y}\mathbf{y}')\mathbf{G}'$ is diagonal. Thus, as noted in Section 4, the Kalman filter can be viewed as an algorithm for exactly diagonalizing the covariance matrix of $\mathbf{y}$.

For ARMA models, an alternative to calculating the exact Gaussian likelihood is to approximate the likelihood by conditioning on the first few y's and ε's. After conditioning, the remaining y's can be written as an invertible linear function of a finite number of current and lagged innovations. Thus, approximating the likelihood by conditioning is equivalent to finding a triangular linear transform of the data having a scalar covariance matrix and is closely related to the linear transform employed by the Kalman filter. More precisely, suppose one used as the initial variance matrix V_1, not the stationary variance given in equation (5), but instead some variance satisfying

$$V_1 = \Sigma + A V_1 A' - \frac{A V_1 b b' V_1 A'}{b' V_1 b + \sigma^2}.$$

Then the iteration scheme (4) produces a constant matrix V_t and the term $b' V_t b + \sigma^2$ appearing in the likelihood (6) does not depend on t. If that term does not depend on the unknown ARMA coefficients either, the Gaussian maximum likelihood estimator minimizes the sum of squared innovations$\sum (y_t - b' a_t)^2$. Thus, using the Kalman filter after setting initial conditions to produce a constant V_t matrix is equivalent to conditioning on initial values and computing nonlinear least squares estimates.

There is still one more statistical procedure that involves a linear transformation approximately diagonalizing the covariance matrix of $\mathbf{y}$. If the y's are a stationary stochastic process, the $T \times T$ Fourier matrix $\mathbf{F}$, with elements $f_{kt} = e^{2\pi i k t / T}$, not depending on any unknown parameters, approximately diagonalizes any stationary covariance matrix. The variable $\mathbf{z} = \mathbf{F}\mathbf{y}$

is the Fourier transform of $\mathbf{y}$ and is the starting point for spectral analysis of time series data. Whereas the variances of the e's are interpreted as forecast error variances (and are constant under the conditioning approach), the variances of the z's (often called the spectrum) are measures of the relative importance of the various cyclical components of the time series.

Although spectral (or frequency domain) analysis can be viewed as a computational device for simplifying the calculation of the parametric Gaussian likelihood function, it is more commonly viewed as a nonparametric approach to studying time series data. It is usually used when the sample size is very large and little structure is imposed except stationarity. Indeed, studying the spectrum using smoothed periodogram values is essentially equivalent to studying the autocorrelation function without assuming a parametric model. In contrast, state space models (e.g., ARMA) impose considerable structure and typically have only a small number of unknown parameters. In addition, stationarity is not necessary. Perhaps because data are so limited and stationarity often implausible, economists seem to prefer the state-space approach to modelling. The Kalman filter is then available as a convenient computational tool.

7 Nonlinear State-Space Models

If we drop the assumption that u_t and v_t are normal, best one-step-ahead predictors are no longer linear in the y's. Maximizing the normal likelihood using the *linear* Kalman filter yields consistent estimates, but at the cost of some efficiency loss. Exact maximum likelihood using a *nonlinear* filter is computationally feasible in low-dimensional problems even if the $\{\alpha_t\}$ process is not autoregressive as long as it is Markovian; that is, as long as the conditional density of α_t given all past α's depends only on α_{t-1}.

Consider the state-space model with measurement equation

$$y_t = b'\alpha_t + u_t$$

where the u_t are i.i.d. with marginal density function $f(\cdot)$. The p-dimensional state vectors $\{\alpha_t\}$ are a Markov process, independent of the process $\{u_t\}$, with joint conditional density

$$\Pr[x \leq \alpha_t \leq x + dx| \text{ all past } \alpha\text{'s}] = h(x|\alpha_{t-1})dx.$$

Again, let Y_t denote the vector $y_1, ..., y_t$. The independence and Markovian assumptions imply that the conditional density of y_t, given Y_{t-1} and α_t, is given by $f(y_t - b'\alpha_t)$ and that the conditional density of α_t, given Y_{t-1} and α_{t-1}, is $h(\alpha_t|\alpha_{t-1})$; that is, they do not depend on past y's.

The likelihood function is the product of the conditional densities $p(y_t|Y_{t-1})$ for $t = 1, ..., T$. If $g(\alpha_t|Y_{t-1})$ is the conditional density of α_t given Y_{t-1}, we have

$$p(y_t|Y_{t-1}) = \int f(y_t - b'\alpha_t)g(\alpha_t|Y_{t-1})d\alpha_t \ . \tag{10}$$

Using Bayes rule for manipulating conditional probabilities, we find

$$\begin{aligned} g(\alpha_t|Y_{t-1}) &= \int h(\alpha_t|\alpha_{t-1})g(\alpha_{t-1}|Y_{t-1})d\alpha_{t-1} = \int h(\alpha_t|\alpha_{t-1})g(\alpha_{t-1}|y_{t-1}, Y_{t-2})d\alpha_{t-1} \\ &= \int h(\alpha_t|\alpha_{t-1})\frac{f(y_{t-1} - b'\alpha_{t-1})g(\alpha_{t-1}|Y_{-2})}{\int f(y_{t-1} - b'\alpha_{t-1})g(\alpha_{t-1}|Y_{t-2})d\alpha_{t-1}}d\alpha_{t-1}. \end{aligned} \tag{11}$$

If f and h are known functions and we have an initial density $g(\alpha_1)$, equation (11) is a recursive relation defining g for period t in terms of its value in period $t-1$. If f and h are normal densities, the integrals are easily evaluated and we find the usual Kalman up-dating formula. Otherwise, numerical integration usually is required.

If α takes on only a finite number of discrete values, g is a mass function and the integration is replace by summation. The calculations then simplify. Suppose α_t is a scalar random variable taking on K different values $r_1, ..., r_K$. Let $\mathbf{g}_t$ be the K-dimensional vector whose k'th element is $g(r_k|Y_{t-1}) \equiv Pr[\alpha_t = r_k|Y_{t-1}]$. Let $\mathbf{H}_t$ be the $K \times K$ Markov matrix whose ij element is $Pr[\alpha_t = r_i|\alpha_{t-1} = r_j]$. Let $\mathbf{f}_t$ be the K-dimensional vector whose k'th element is $f(y_t - br_k)$ and let $\mathbf{z}_t$ be the K-dimensional vector whose k'th element is $f_{tk}g_{tk}$. The likelihood function is

$$\prod_{t=1}^{T} p(y_t|Y_{t-1}) = \prod_{t=1}^{T} \mathbf{f}_t'\mathbf{g}_t$$

where, from (11), the g's can be computed from the recursion

$$\mathbf{g}_t = \frac{\mathbf{H}_t\mathbf{z}_{t-1}}{\mathbf{f}_{t-1}'\mathbf{g}_{t-1}}$$

A simple example is Hamilton's *Markov switching model.* We assume

$$y_t = \beta' x_t + \delta' x_t \alpha_t + u_t$$

where α_t is a binary zero-one Markovian random variable such that $Pr[\alpha_t = 1|\alpha_{t-1} = 1] = p$ and $Pr[\alpha_t = 0|\alpha_{t-1} = 0] = q$. Thus, with probability $1 - q$ we switch from a regime where $E[y_t] = \beta' x_t$ to a regime where $E[y_t] = (\beta + \delta)' x_t$; we switch back with probability $1 - p$.