Chapter 7

Interpolation and Distribution

One important (and non-trivial) use of state-space techniques for single series is interpolation or distribution of data. These techniques are often known generically as "interpolation", but, as pointed out by Chow and Lin, there are actually two quite distinct applications: *interpolation*, which is applied (primarily) to stocks, and *distribution*, which is applied to flows. By far the more commonly used of these is distribution, in which the constructed series is designed to sum to the observed series over the course of the interval. We'll concentrate on the latter, which is much more difficult.

It's assumed that we may have some useful information from other data at the target frequency. The following set of options takes in almost all known techniques for this. The connection between the high-frequency information and the interpolated series is represented by one of the models:

$$I_t = H_t + u_t \tag{7.1}$$

$$I_t = H_t u_t \tag{7.2}$$

$$\log\, I_t = H_t + u_t \tag{7.3}$$

The high-frequency information from related series takes one of the forms:

$$H_t = O_t\beta, O_t \text{ observable} \tag{7.4}$$

$$H_t \text{ observable} \tag{7.5}$$

$$H_t = 0 \tag{7.6}$$

and the time series model for the deviation between the high-frequency and the interpolated data is one of:

$$u_t = \rho u_{t-1} + v_t \tag{7.7}$$

$$u_t = u_{t-1} + v_t \tag{7.8}$$

$$u_t = (1+\rho)u_{t-1} - \rho u_{t-2} + v_t \tag{7.9}$$

While it's possible to allow for a more complicated time series structure for the noise, in practice, there's no need for that; the requirement that the data hit known values at a regular interval will dominate any short-term dynamics.

The observation equation connecting the interpolated data with the observed

data is one of:

$$L_t = I_t, t = q, 2q, \ldots \tag{7.10}$$

$$L_t = \sum_{s=0}^{q-1} I_{t-s}, t = q, 2q, \ldots \tag{7.11}$$

where q is the ratio between the desired higher frequency and the observed lower frequency: $3 = 12/4$ for quarterly to monthly. The first of these is the formal definition of interpolation, while the second is distribution.

7.1 Linear Model

The combination of (7.1) with (7.4) and (7.7) forms the procedure from Chow and Lin (1971). While there are many examples of the use of this, it seems likely that this is misspecified. If (as is typical), I_t is an $I(1)$ process, then it has to be co-integrated with some member of O_t in order for the difference to be stationary. This is probably overstating that relationship. Replacing (7.7) with (7.8) is the method from Fernandez (1981), while using (7.9) is from Litterman (1983).

While including a unit root in the disturbance process is likely to be much closer to a realistic description of a relationship than the stationary disturbance in Chow-Lin,[1] the linear relationship between a variable like GDP and its related series would again seem misspecified; such variables are almost always modeled in logarithms. The choice of the linear relationship (7.1) is more a matter of computational convenience when combined with the adding-up constraint (7.11). After showing how to handle the strictly linear models with state-space techniques, we'll move on to the more realistic log-linear model.

A couple of things to note. First, all the combinations above can be done using the `DISAGGREGATE` procedure, so you don't have to do any programming if you just want to employ these methods in their standard forms. Second, if you read any of the papers cited above, you'll note that none of them (explicitly) use state-space techniques. There are many published papers which, in effect, rederive Kalman smoothing calculations for specific models (or just invert a full $T \times T$ matrix when Kalman smoothing could do the calculation more efficiently).

In translating the linear models into state-space form, the states will be u_t and its lags. How many lags to we need? The evolution of the states requires none for (7.7) and (7.8) and one for (7.9).[2] Where we will need extra lags is in the (final form of) the measurement equation (7.11). We'll need q total for that, so we'll have to expand the state vector with lags to give us $u_t, u_{t-1}, \ldots, u_{t-q+1}$.

If we substitute (7.1) into (7.11), we get a measurement equation of:

$$L_t = (O_t + \ldots + O_{t-q+1})\,\beta + [1, \ldots, 1]\,[u_t, \ldots, u_{t-q+1}]' \tag{7.12}$$

[1] Litterman shows that it produces more accurate values for a number of examples.

[2] Remember that the first lag will be in the lag term in the state equation.

Note that we only have an observation on this every q periods. $O_t + \ldots + O_{t-q+1}$ is known, so given β, we can subtract those terms off the observable to get:

$$L_t - (O_t + \ldots + O_{t-q+1})\,\beta = [1, \ldots, 1]\,[u_t, \ldots, u_{t-q+1}]' \tag{7.13}$$

The combination of (7.13) and the appropriate state representation for the u_t process will form a state-space model for the data. We can estimate the β by maximum likelihood and use Kalman smoothing to get the distributed series I_t. Kalman smoothing gives us smoothed estimates of u_t; we just need to add that to the computed values of H_t.

One thing to note is that the noise model (7.7) is stationary, (7.8) has one unit root and (7.9) has one unit root and one stationary root. You can use the option `PRESAMPLE=ERGODIC` to handle the initial mean and variance for any of these. If you look at the `DISAGGREGATE` procedure, it uses a `G` matrix which is a set of difference operations for the two unit root choices. Either method is fine. The stationary noise model has a zero mean, so the O_t variables should include a constant. The unit root noise processes can seek their own level, so O_t should *not* include a constant; it will be redundant.

7.2 Log-Linear Model

Let's now look at how we can handle the more realistic log-linear relationship. (7.11) would now need to be written:

$$L_t = \exp(\log\, I_t) + \ldots + \exp(\log\, I_{t-q+1})$$

since our state-space model is able to generate $\log I_t$. The state equation is linear in $\log I_t$, but the measurement equation is non-linear. What we will describe now is the simplest case of the ***Extended Kalman Filter***, which generalizes the Kalman filter to allow for non-linearities. This treats the non-linearities by repeated linearizations. Each linearized version can be solved using the standard Kalman filter.

About what do we linearize? $\log I_t$ is a function of the O_t variables and the unobservable states u_t. We can use the values for that from a previous linearization pass (for iteration 1, taking $u_t = 0$). For convenience, write $i_t = \log I_t$, and $\widetilde{i}_t$ as the Kalman smoothed estimate of i_t from the previous iteration. Then the linearized version is:

$$\begin{aligned} L_t \approx\; & \exp(\tilde{i}_t) + \ldots + \exp(\tilde{i}_{t-q+1}) \\ & + \exp(\tilde{i}_t)(i_t - \tilde{i}_t) + \ldots + \exp(\tilde{i}_{t-q+1})(i_{t-q+1} - \tilde{i}_{t-q+1}) \\ & = \exp(\tilde{i}_t)\left(1 - \tilde{i}_t\right) + \ldots + \exp(\tilde{i}_{t-q+1})\left(1 - \tilde{i}_{t-q+1}\right) \\ & + \exp(\tilde{i}_t) i_t + \ldots + \exp(\tilde{i}_{t-q+1}) i_{t-q+1} \end{aligned} \tag{7.14}$$

The first line terms in (7.14) are fixed for a given iteration and can be subtracted from L_t. The related regressors are a bit more complicated now, since,

when we substitute in the definition of i_t, we get for the second set of terms:

$$\left(\exp(\tilde{i}_t)O_t + \ldots + \exp(\tilde{i}_{t-q+1})O_{t-q+1}\right)\beta + \exp(\tilde{i}_t)u_t + \ldots + \exp(\tilde{i}_{t-q+1})u_{t-q+1}$$

The combined "regressor" to determine β now changes from iteration to iteration, as do the weights on the states.

There's one minor problem with how this will operate. Iterating on the above gives us a solution for the states given β. But β needs to be estimated as well. One possibility is to update the expansion point with each iteration on β, that is, do just one linearization per iteration. However, there's a simpler way to handle this. Given the rest of the structure of the state-space model, the model is a linear equation in β, which can be estimated directly by a form of GLS. In fact, the papers cited all show the form of the GLS estimator. With **DLM**, the most convenient way to do this is to add the coefficients to the states, glueing together a model with the noise model states plus

$$\beta_t = \beta_{t-1} + 0$$

The loadings on those "states" will be

$$\left(\exp(\tilde{i}_t)O_t + \ldots + \exp(\tilde{i}_{t-q+1})O_{t-q+1}\right)$$

When you do this, the coefficients should be the last block of states, and you should include the option FREE=*number of regressors*. That adjusts the log likelihood so it will give the value conditional on β, which is the form you would use in doing this by GLS.

7.3 Proportional Denton Method

As an example, we'll look a distribution method which uses just a single related series. This is the proportional Denton method. You can find a description of this in Bloem, Dippelsman, and Maehle (2001), chapter 6. Statistics bureaus often have different reads on the same series at different recording frequencies. The general assumption is that the coarser the collection interval, the more accurate the data. For instance, in the U.S., some of the information which goes into GDP calculations comes from tax returns, which are only available well into the next year, while other information is available monthly or quarterly. The proportional Denton method assumes that (for instance), the annual data are accurate. The distributed quarterly data are to sum to the annual value, but should (roughly) show the movements apparent in the observations of the less accurate observed quarterly data.

The proportional Denton method is described on page 87 of the IMF manual as the solution to [3]:

$$\min \sum_{t=2}^{T} \left(\frac{I_t}{H_t} - \frac{I_{t-1}}{H_{t-1}} \right)^2$$

[3]Variable names have been changed to match our discussion

over $\{I_t\}$ subject to

$$L_t = I_t + \ldots + I_{t-q+1}; t = q, 2q, \ldots$$

You'll notice again that this is not written in state-space form. However, if we define u_t by $I_t = H_t u_t$, then, since H_t are known, the problem can be rewritten as:

$$\min \sum_{t=2}^{T} (u_t - u_{t-1})^2 \tag{7.15}$$

over $\{u_t\}$ subject to

$$L_t = [H_t, \ldots, H_{t-q+1}][u_t, \ldots, u_{t-q+1}]'; t = q, 2q, \ldots \tag{7.16}$$

The solution to this turns out to be the same as Kalman smoothing on the model with state equation (augmented by lags):

$$u_t = u_{t-1} + w_t$$

and measurement equation (7.16). The sum in (7.15) is just the sum of squared w_t. Since those are assumed to be i.i.d. Normal, that, in effect, gives the unconditional density for the w_t. Kalman smoothing gives us the *conditional* means of the states and the disturbances subject to the observations.

Example **7.1** does an annual to quarterly distribution so $q = 4$. This means the A and F matrices can be set up by:

```
dec rect a(4,4)
input a
 1 0 0 0
 1 0 0 0
 0 1 0 0
 0 0 1 0
compute f=%unitv(4,1)
```

As we can see from (7.16), the C matrix here depends upon the data. We can set this up with

```
dec frml[vec] cf
frml cf = ||quarter{0},quarter{1},quarter{2},quarter{3}||
```

The only thing not "known" is the variance of w_t, but since we're interested only in the means, we don't even need to know that or estimate it. We can just make it 1.0 and be done. The distributed values are obtained here by multiplying the observed quarterly data by the Kalman smoothed estimates of u_t:

```
dlm(type=smoothed,a=a,c=cf,y=annual,f=f,sw=1.0,$
  presample=ergodic) 1998:1 2000:4 xstates
set distrib = %scalar(xstates)*quarter
```