Ragged-data Models

kassandra

August 30, 2018

1 Massimiliano Marcellino, Christian Schumacher: "Factor-MIDAS for Now- And Forecasting with Ragged-Edge Data: A Model Comparison for German GDP" (2008)

They focus on a quarterly GDP growth (y_{t_q}), where t_q is a quarter time index ($t_q = 1, 2, 3, \ldots, T_q$). Also they consider monthly frequency (y_{t_m}): $y_{t_m} = y_{t_q} \forall t_m = 3t_q \Rightarrow t_m = 3, 6, 9, \ldots, T_m, T_m = 3T_q$.

Their aim is to predict GDP growth for h_q quarters ($h_m = 3h_q$ months) ahead. They do it by:

1. Taking the factors out of data:

$$X_{t_m} = \Lambda F_{t_m} + \xi_{t_m},$$

where

X_{t_m} is an information set, N-dimensional vector containing a large set of stationary monthly indicators,

Λ is a loadings matrix ($N \times r$),

F_{t_m} is a r-dimensional factor vector,

ξ_{t_m} is a component of X_{t_m} not explained by factors.

2. Applying the MIDAS model to these factors.

They present three ways to take the factors out.

1.1 Way 1: Vertical Realignment of Data and Dynamic Principal Components Factors (Based on Altissimo et al. (2006))

The idea is to realign each time series in a sample to obtain a balanced dataset.

Assume variable i has a publication lag of K_i months. Then for period T_m the last available observation will be $T_m - K_i$. The realignment procedure is the following one:

$$\tilde{x}_{i,T_m} = x_{i,T_m - K_i}$$

for $t_m = K_i + 1, \ldots, T_m$.

Applying this procedure to all the series and harmonizing the beginning of the sample, we get a balanced dataset $\tilde{X}_{t_m}$ for $t_m = \max(\{k_i\}_{i=1}^N) + 1, \ldots, T_m$.

Then dynamic PCA is applied to estimate factors.

Positive: it is simple. Negative: statistics is published at different moments and factors may be reassessed more often than the model frequency is. These lead to realignment changes the correlation structure of data (however, there is an opinion stated that dynamic PCA is able to overcome this problem).

1.2 Way 2: Principal Components Factors and the EM algorithm (Basrd on Stock and Watson (2002))

Assume variable i from X_{t_m} is presented with a full vector $X_i = (x_{i,1} \ldots x_{i,T_m})$. If there is a ragged-edge problem, let a vector X_i^{obs} contain all the available observations ($X_i^{obs} \subset X_i$):

$$X_i^{obs} = A_i X_i,$$

where A_i is a matrix which can tackle missing values or frequencies. If there is no missing data, A_i is an identity matrix. If there is missing data in the end of the sample (due to the rag), the corresponding row is deleted from A_i. Then the EM algorithm is applied:

1. Make a naive prediction of $\hat{X}_i^{(0)} \forall i \Rightarrow$ get a balanced dataset $\hat{X}^{(0)}$. Standard PCA provides $\hat{F}^{(0)}$ and $\hat{\Lambda}^{(0)}$.

2. E-step. Update:

$$\hat{X_i}^{(j)} = \hat{F}^{(j-1)}\hat{\Lambda}_i^{(j-1)} + A_i'(A_i'A_i)^{-1}(X_i^{obs} - A_i\hat{F}^{(j-1)}\hat{\Lambda}_i^{(j-1)}).$$

3. M-Step. Repeat E-step for all i, getting balanced dataset. Reestimate $\hat{F}^{(j)}$ and $\hat{\Lambda}^{(j)}$ using PCA, repeat E-Step until convergence.

1.3 Way 3: Estimation of a Large Parametric Factor Model in The State-Space Form (Based on Doz et al. (2006))

An explicit dynamic VAR structure is assumed to hold for the factors. The full state-space model is as following:

$$X_{t_m} = \Lambda F_{t_m} + \xi_{t_m}$$

$$\Psi(L_m)F_{t_m} = B\eta_{t_m}$$

The second row is the VAR of the factors, where:

$\Psi(L_m) = \sum_{i=1}^{p} \Psi_i L_m^i$ is a lag polynomial,

L_m is a monthly lag operator, $L_m x_{t_m} = x_{t_m-1}$,

η_{t_m} is a q-dimensional vector containing orthogonal dynamic shocks that drive r factors.

B is a $(r \times q)$ matrix.

If the dimension of X_{t_m} is small, the model is estimated using ML. Otherwise, using quasi-ML:

1. Estimate $\hat{F_{t_m}}$ using PCA as an initial estimate.

2. Estimate $\hat{\Lambda}$ by regressing X_{t_m} on the estimated factors $\hat{F_{t_m}}$. Estimate covariance $\hat{\xi}_{t_m} = X_{t_m} - \hat{\Lambda}\hat{F}_{t_m}$ (this covariance is denoted as $\hat{\Sigma}_\xi$).

3. Estimate factor VAR(p) on the factors $\hat{F}_{t_m}$ which gives $\hat{\Psi}(L)$ and the residual covariance of $\hat{c}_{t_m} = \hat{\Psi}(L_m)\hat{F}_{t_m}$. This covariance is denoted as $\hat{\Sigma}_c$).

4. To obtain B with given q, perform eigenvalue decomposition of $\hat{\Sigma}_c$). Let M be an $(r \times q)$ matrix containing eigenvectors corresponding to q biggest eigenvalues, and P be a $(q \times q)$ matrix, containing these biggest eigenvalues on the main diagonal and zeros otherwise. Then $\hat{B} = \frac{M}{\sqrt{P}}$.

5. The coefficients and auxiliary parameters are fully specified numerically. The model is cast into space-state form. To estimate factors use Kalman filter or smoother.

1.4 Factor-MIDAS: Predict Monthly GDP Growth Using Estimated Factors

MIDAS stands for "mixed-data sampling".

1.4.1 Basic Factor-MIDAS

General model:

$$y_{t_q+h_q} = y_{t_m+h_m} = \beta_0 + \sum_{i=1}^{r} \beta_{1,i} b_i(L_m, \theta_i)\hat{f}_{i,t_m}^{(3)} + \varepsilon_{t_m+h_m}.$$

Considering a case with $r = 1$, i.e. only one factor $\hat{f}_{t_m}$ is used, the forecast for h_q quarters ($h_m = 3h_q$ months) is as following:

$$y_{t_q+h_q} = y_{t_m+h_m} = \beta_0 + \beta_1 b(L_m, \theta)\hat{f}_{t_m}^{(3)} + \varepsilon_{t_m+h_m},$$

where

$b(L_m, \theta) = \sum_{k=0}^{K} z(k, \theta)L_m^k$

$z(k, \theta) = \frac{\exp(\theta_1 k + \theta_2 k^2)}{\sum_{k=0}^{K} \exp(\theta_1 k + \theta_2 k^2)}$

$\hat{f}_{t_m}^{(3)} = \hat{f}_{t_m} \forall t_m = \ldots, T_{m-6}, T_{m-3}, T_m$. Lags are treated accordingly.

The model can be estimated using nonlinear least squares in a regression in a regression of y_{t_m} onto $\hat{f}_{t_m-k}^{(3)}$, which gives coefficients $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\beta}_0$, $\hat{\beta}_1$.

The forecast is as following:

$$y_{T_m+h_m|T_m} = \hat{\beta}_0 + \hat{\beta}_1 b(L_m, \hat{\theta})\hat{f}_{T_m}.$$

1.4.2 Smoothed MIDAS (Based on Altissimo et al. (2006))

They consider New Eurocoin index and present the following projection:

$$y_{T_m+h_m|T_m} = \hat{\mu} + G\hat{F}_{T_m}$$

$$G = \tilde{\Sigma}_{yF}(h_m) \times \hat{\Sigma}_F^{-1},$$

where

G is the projection coefficient matrix.

$\tilde{\Sigma}_{yF}(h_m)$ is the cross-covariance with h_m monthly lags between the GDP growth and factors.

$\hat{\Sigma}_F$ is the estimated sample covariance of the factors.

The smooth component is within $\tilde{\Sigma}_{yF}(h_m)$. Assume the factors and the GDP growth are demeaned. Then let the covariance between $\hat{F}_{t_m-k}$ and y_{t_m} be estimated by

$$\hat{\Sigma}_{yF}(k) = \frac{1}{T^* - 1} \sum_{t_m=M+1}^{T_m} y_{t_m} \hat{F}^{(3)'}_{t_m-k},$$

where $T^* = \text{floor}[(T_m - (M+1))/3]$ is the number of observations available to compute cross-covariance for $k = -M \dots M$ and $M \geq 3h_q = h_m$.

Then the estimation of cross-spectral matrix is as following:

$$\hat{S}_{yF}(w_j) = \sum_{k=-M}^{M} \left(1 - \frac{|k|}{M+1}\right) \hat{\Sigma}_{yF}(k) e^{iw_j k}$$

at frequencies $w_j = \frac{2\pi j}{2H}$ for $i = -H, \dots, H$ (Bartlett lag window). By reverse Fourier transform they obtain:

$$\tilde{\Sigma}_{yF}(k) = \frac{1}{2H+1} \sum_{j=-H}^{H} \alpha(w_j) \hat{S}_{yF}(w_j) e^{iw_j k}$$

where $\alpha(w_j)$ is a frequency-response function. In Eurocoin example,

$$\alpha = \begin{cases} 1, & \forall |w_j| < \frac{\pi}{6} \\ 0, & \text{otherwise.} \end{cases}$$

1.4.3 The Unrestricted MIDAS (Based on Marcellino and Schumacher (2007))

They consider an unrestricted lag order model:

$$y_{T_m+h_m} = \beta_0 + D(L_m) F^{(3)}_{t_m} + \varepsilon_{t_m+h_m},$$

where $D(L_m) = \sum_{k=0}^{K} D_k L_m^k$ is an unrestricted lag polynomial of frequency k. $D(L_m)$ and β_0 are estimated by OLS.

2 Kees E. Bouwman, Jan P.A.M. Jacobs: "Forecasting with real-time macroeconomic data: the ragged-edge problem and revisions" (2005)

They claim that considering linear time series models ragged-edge data problem can be solved by the Kalman filter (Harvey (1989) and Hamilton (1994, Chapter 13)). They assume the following:

1. The maximum publication lag is equal to one month.

2. $x_1(t), t \in N$ is the vector of final values for period t of the variables released without a publication lag.

3. $x_2(t), t \in N$ is the vector of final values for period t of the variables released with an one-month publication lag.

4. The data becomes final by the end of the 5-month period since the first publication, so that there is a maximum of 5 data releases and 4 data revisions (updates) for a period. $x_k(i,t), k = 1, 2, i = 1, \dots, 5$ is the i-th release of the value x_k for a period t. The release period of $x_k(i,t)$ is denoted by $\tau_k(i,t)$:

$$\tau_1(i,t) = t + i,$$

$$\tau_2(i,t) = t + i + 1.$$

   As the 5-th release is final, $x_k(t) \equiv x_k(5,t)$ and $\tau_k(t) \equiv \tau_k(5,t)$.

5. A new vintage is released every month. All values in a vintage are given by their latest available release:

$$x_1^T(t) = x(\min(T - t, 5), t) \forall t < T,$$

$$x_2^T(t) = x(\min(T - t - 1, 5), t) \forall t < T - 1,$$

   where T is the vintage date.

6. ...