Inference for Factor-MIDAS Regression Models

Yookyung Julia Koh

McGill University

Introduction

- **MIDAS(MIxed-DAta Sampling)** regression is a popular forecasting/nowcasting mixed-frequency estimation method.

  e.g. forecast/nowcast quarterly GDP growth rate with monthly indicators.

- **Factor model**: instead of using many predictors, we can use a few common factor from a large panel of high-frequency predictors.

⇒ **Factor-augmented MIDAS (factor-MIDAS) regression models".**

- Our objective is to study the **inference methods for factor-MIDAS regression models**.

Motivation

- The factors are latent, hence the estimation in factor-MIDAS proceeds in two steps.
  1. Estimate the factors by principal component analysis (PCA).
  2. Estimate the parameters in the regression model.
- This two step estimation procedure complicates the estimation.
- Previous literature show that there exists an asymptotic bias in the factor-augmented regression models due to the two step estimation procedure (e.g., [Ludvigson and Ng(2010)] and [Gonçalves and Perron(2014)]).

Contribution

- We derive the **asymptotic distribution** of the estimators in the factor-MIDAS regression models.
- We find that there exists an **asymptotic bias**.
- The bias depends on the **serial dependence** and **cross-sectional dependence** in the idiosyncratic errors of the factor model.
- We propose two **bias correction methods**:
  - **analytical bias correction** based on asymptotic theory, and
  - **a bootstrap method**.
- Both simulation and empirical results illustrate well that there is a significant bias, indicating the importance of correcting it.

The model

Factor-MIDAS regression model:

$$y_t = \beta_0 + \beta_1' W(L^{1/m};\theta) f_t + \varepsilon_t, \quad t = 1,\dots,T$$

- $W(L^{1/m};\theta) f_t = w_0(\theta) + w_1(\theta) f_{t-1/m} + \dots + w_K(\theta) f_{t-K/m}$.
- f_t is a $r \times 1$ vector of common factors in the following high-frequency panel factor model,

  $$X_{t-k/m} = \Lambda f_{t-k/m} + e_{t-k/m},$$

  for $k = m-1,\dots,0$ and $t = 1,\dots,T$. (High-frequency variables are observed at most m times between t and $t-1$.)
- $w_k(\theta) = diag(w_{k,1}(\theta_1),\dots,w_{k,r}(\theta_r))$ is a $r \times r$ diagonal matrix of weighting scheme for k-th lag of factors, $k = 1,\dots,K$.
- Common weighting scheme: exponential Almon lag with two parameters such that $w_{k,j}(\theta_j) = \frac{\exp(\theta_{j,1}k + \theta_{j,2}k^2)}{\sum_{k=0}^{K}\exp(\theta_{j,1}k+\theta_{j,2}k^2)}, j = 1,\dots,r$.
- Two-step estimation procedure:
  1. Estimate the factors by PCA ⇒ $\tilde{f}_{t-k/m}$.
  2. Estimate $\beta = (\beta_0, \beta_1')'$ and $\theta = (\theta_1, \theta_2)'$ by nonlinear least squares ⇒ $\tilde{\beta} = (\tilde{\beta}_0, \tilde{\beta}_1')'$, $\tilde{\theta} = (\tilde{\theta}_1, \tilde{\theta}_2)'$.

Asymptotic distribution

Using the estimated factors, if $\sqrt{T}/N \to c$, $0 \le c < \infty$,

$$\sqrt{T}(\tilde{\alpha} - \alpha) \xrightarrow{d} N(-c\Delta_\alpha, \Sigma_\alpha),$$

where $\tilde{\alpha} = (\tilde{\beta}', \tilde{\theta}')'$ is NLS estimator and $-c\Delta_\alpha = -c(\Delta_{\beta_0}, \Delta_{\beta_1}', \Delta_\theta')'$.

$\Delta_{\beta_0} = 0$ and the bias, $-c\Delta_{\beta_1}$ and $-c\Delta_\theta$ depend on:

- c, which is the limit of $\sqrt{T}/N$,
- the variance of the factor estimation error, which is a function of $Var\left(\frac{\Lambda' e_{t-k/m}}{\sqrt{N}}\right)$, and
- the covariance of the scaled average of the factor loadings and idiosyncratic errors between distinct time, $Cov\left(\frac{\Lambda' e_{t-k/m}}{\sqrt{N}}, \frac{\Lambda' e_{t-l/m}}{\sqrt{N}}\right)$, for $k \neq l$.

Therefore, the asymptotic bias depends on both serial and cross sectional dependence in the idiosyncratic errors of the factor model.

Inference method: plug-in bias correction

- We replace the unknown parameters by consistent estimators.
- The key is to consistently estimate a term in the bias, $\Gamma_k = Cov(\frac{\Lambda' e_t}{\sqrt{N}}, \frac{\Lambda' e_{t-k/m}}{\sqrt{N}})$.
- We propose a consistent estimator for Γ_k allowing for serial and cross-sectional dependence.
- $\sqrt{T}(\tilde{\alpha}_{\text{BC}} - \alpha) \xrightarrow{d} N(0, \Sigma_\alpha)$, where $\tilde{\alpha}_{\text{BC}}$ is bias corrected estimator, $\tilde{\alpha}_{\text{BC}} = \tilde{\alpha} - (-\tilde{\Delta}_\alpha)$.

Alternative inference method: a bootstrap method

- We propose a two-step bootstrap procedure:
  1. resample the factor model and the nonlinear regression model by a residual-based bootstrap, and then
  2. obtain bootstrap nonlinear estimates.
- We propose a new bootstrap method for factor model that is robust to serial and cross-sectional dependence, which we call AR-sieve + CSD bootstrap and show that this method is asymptotically valid.
- AR-sieve + CSD bootstrap is an application of the autoregressive sieve bootstrap to the idiosyncratic residuals of each time series in the panel data and the innovations are resampled using cross-sectional dependent bootstrap, proposed by [Gonçalves and Perron(2020)].

Monte Carlo Simulation

	$N=50$			$N=100$			$N=200$		
	$T=50$	100	200	50	100	200	50	100	200
bias of β_1									
True Factor	0.00	0.00	0.00	-0.01	0.00	0.00	-0.01	0.00	0.00
Estimated Factor	-0.64	-0.57	-0.54	-0.41	-0.35	-0.31	-0.28	-0.21	-0.18
Plug-in	-0.45	-0.42	-0.41	-0.26	-0.26	-0.25	-0.14	-0.14	-0.14
AR-sieve+CSD	-0.23	-0.23	-0.24	-0.17	-0.16	-0.16	-0.12	-0.10	-0.10
95% coverage rate of β_1									
Estimated Factor	52.2	44.5	29.2	72.3	71.8	67.3	81.5	85.0	84.1
Plug-in	72.0	77.1	77.1	81.1	86.0	87.9	85.0	90.1	91.3
AR-sieve+CSD	86.3	80.0	73.5	91.0	89.8	87.1	93.2	93.2	92.6

The table shows the simulation results (size of the bias and the coverage rates) of the factor-MIDAS regression model with a single factor. T is the time series dimension of the low-frequency target variable. N denotes the cross-sectional dimension in the high-frequency factor model. The error terms in the regression model are generated by GARCH model to allow for heteroskedasticity: $\varepsilon_t = \sqrt{h_t} v_t$, where $h_t = 0.1 + 0.3\varepsilon_{t-1}^2 + 0.6 h_{t-1}$ and $v_t \sim$ i.i.d.$N(0,1)$. The idiosyncratic error terms are generated by AR(1) for each series: $e_{i,t-k/m} = a_i e_{i,t-(k-1)/m} + u_{i,t-k/m}$, where $a = a_i = 0.5$ and $corr(u_{i,t-k/m}, u_{j,t-k/m}) = 0.5^{|i-j|}$ if $|i-j| \le 5$ and 0 for otherwise. Note that the coverage rate results of "estimated factors" and "plug-in" are based on asymptotic theory. The bootstrap coverage rates use the bootstrap equal-tailed percentile t method.

Empirical application

We nowcast a quarterly U.S. real GDP growth rate with monthly macroeconomic factors: for each $h = 2, 1, 0$,

$$y_t = \beta_0 + \beta_1' W(L^{1/m};\theta) f_{t-h/m} + \rho y_{t-1} + \varepsilon_t.$$

- Sample period: 1984 Q4 - 2022 Q4.
- h represents a nowcasting horizon. (e.g. $h = 2$ indicates that we nowcast a current output growth at 2 months ahead.)
- We obtain two monthly factors from FRED-MD dataset (74 macroeconomic indicators): $f_{1,t-k/m}$ and $f_{2,t-k/m}$ explain 28% and 10% of total variation of the dataset.
- We report 90% confidence interval of the point estimates: based on asymptotic theory and based on bootstrap method.

	$h=2$		$h=1$		$h=0$	
constant	0.90		0.83		0.99	
Asymptotic	0.67	1.01	0.67	0.99	0.78	1.21
AR sieve+CSD	0.71	0.98	0.69	0.94	0.75	1.46
aggregated first factor	2.54		3.79		1.87	
Asymptotic	1.64	3.44	2.97	4.61	0.31	3.44
AR sieve+CSD	2.13	3.54	3.34	4.80	0.90	3.39
aggregated second factor	0.04		0.36		-0.95	
Asymptotic	-0.22	0.30	0.08	0.65	-1.47	-0.43
AR sieve+CSD	-0.12	0.38	0.16	0.77	-1.63	-0.21
y_{t-1}	-0.30		-0.30		-0.58	
Asymptotic	-0.54	-0.06	-0.52	-0.09	-0.87	-0.28
AR sieve+CSD	-0.49	-0.12	-0.43	-0.14	-1.22	-0.25

The first line in each panel gives the point estimates. The row "Asymptotic" presents the confidence interval based on asymptotic theory, by adding and subtracting 1.645 times the heteroskedasticity robust standard errors. The row "AR-sieve+CSD" bootstrap shows the results based on our proposed bootstrap method. We use equal-tailed bootstrap interval with bootstrap number 4999.

Conclusion

- We derive the asymptotic distribution of the estimators in the factor-MIDAS regression models.
- We show that there exists a bias when the cross-section dimension is relatively small to the time series dimension.
- The bias depends on time series as well as cross sectional dependence in the idiosyncratic error term by the fact that the factors and their lags are aggregated.
- We propose and theoretically justify two inference methods for factor-MIDAS regression models: plug-in bias estimator and a bootstrap method.

References

[Gonçalves and Perron(2014)] Sílvia Gonçalves and Benoit Perron. Bootstrapping factor-augmented regression models. *Journal of Econometrics*, 182(1):156–173, 2014.

[Gonçalves and Perron(2020)] Sílvia Gonçalves and Benoit Perron. Bootstrapping factor models with cross sectional dependence. *Journal of Econometrics*, 218(2):476–495, 2020.

[Ludvigson and Ng(2010)] S Ludvigson and S Ng. A factor analysis of bond risk premia, handbook of empirical economics and finance, ed by aman ulah and david a. giles, 313–372, 2010.

McGill UNIVERSITY