Forecasting Using Cross-Section Average-Augmented Time Series Regressions

Joakim Westerlund   Hande Karabiyik

Lund University

Vrije Universiteit Amsterdam

December 16, 2019

The problem of interest

- We want to forecast the time series variable y_t $(t = 1, ..., T)$.
- The model for y_t that we consider is given by

$$y_{t+h} = \alpha' F_t + \beta' W_t + \varepsilon_{t+h} = \delta' z_t + \varepsilon_{t+h}$$

where $z_t = [F_t', W_t']'$ is $(r+n) \times 1$ and $\delta = [\alpha', \beta']'$.

- Problem: F_t is unobserved and potentially correlated with W_t!
- Solution: We assume the existence of an $m \times 1$ panel data variable $x_{i,t}$ $(i = 1, ..., N)$ that loads on the same set of factors as y_t;

$$x_{i,t} = \lambda_i' F_t + e_{i,t}$$

The problem of interest

- We want to estimate the factors from $x_{i,t}$ and use these in place of F_t when forecasting y_t.

- The mean-square optimal forecast is given by

$$y_{T+h|T} = E(y_{T+h}|z_T, z_{T-1}, ...) = \delta' z_T$$

- The feasible forecast is

$$\widehat{y}_{T+h|T} = \widehat{\delta}' \widehat{z}_T$$

where $\widehat{z}_t = [\widehat{F}'_t, W'_t]'$, $\widehat{F}_t$ is the estimated factor and $\widehat{\delta} = [\widehat{\alpha}', \widehat{\beta}']'$ is the OLS slope estimator in a regression of y_{t+h} onto $\widehat{z}_t$.

The problem of interest

- This type of factors-based forecasting has attracted A LOT of attention!
- A few references: Stock and Watson (JASA and JBES 2002), Bai and Ng (ETCA 2006, JE 2008 and JAE 2009), Boivina and Ng (JE 2006), Cheng and Hansen (JE 2015), Choi (ET 2012), Corradi and Swanson (JE 2014), Djogbenou et al. (JTSA 2015 and JBES 2017), Gonçalves and Perron (JE 2014), and Gonçalves et al. (JE 2017).
- Reason: "Both the leading indicator and VAR models perform slightly better than the univariate AR in this simulated out-of-sample experiment. However, the gains are not large. The factor models offer substantial improvement" (Stock and Watson, JASA 2002)

The problem of interest

Figure: Table 2 Stock and Watson (JASA 2002).

Table 2. Simulated Out-of-Sample Forecasting Results Industrial Production, 12-Month Horizon

Forecast method	*Relative MSE*
Univariate autoregression	1.00
Vector autogression	.97
Leading indicators	.86
Principal components	.58
Principal components, $k = 1$	.94
Principal components, $k = 2$	.62
Principal components, $k = 3$	.55
Principal components, $k = 4$	.56
Principal components, AR	.69
Root MSE, AR model	.049

Motivation

- The existing literature is based almost exclusively on using principal components (PC) to estimate F_t.
- In the present paper we use the cross-section average (CA) of $x_{i,t}$ as an estimator of F_t.
- Rationales:
  - Super simple!
  - Intuitive, as we want to forecast the conditional mean.
  - Natural given the good performance of the simple average in forecast combination and interactive effects panel data models.
  - Facilitates easy interpretation of the estimated factors.

This paper

- We consider the asymptotic and small-sample properties of $\widehat{y}_{T+h|T}$ when

$$\widehat{F}_t = \overline{x}_t = \frac{1}{N}\sum_{i=1}^{N} x_{i,t}$$

- We do what Bai and Ng (ETCA 2006) do for PC under the same conditions, except that we
  - allow $r \leq m$ to be unknown,
  - need $m \geq 1$ panel variables, and
  - require $\mathrm{rk}\,\overline{\lambda} = r \leq m$.

Assumptions

- $e_{i,t}$ is mean zero, but may be heteroskedastic and weakly dependent across both i and t.
- λ_i, F_t and $e_{i,t}$ are independent, and z_t and ε_t are independent of $e_{i,t}$.
- $E(\varepsilon_{t+h}|z_t, z_{t-1}, ...) = 0$ for $h > 0$.
- z_t may be weakly dependent and can include y_t.
- $\text{plim}_{T\to\infty} T^{-1} \sum_{t=1}^{T} z_t z_t'$ is positive definite.
- $\text{rk}\,\overline{\lambda} = r \leq m$ for all N, including $N \to \infty$.

Asymptotics

- We can show that (under $r = m$)

$$\widehat{y}_{T+h|T} - y_{T+h|T} = T^{-1/2}\sqrt{T}(\widehat{\delta} - \delta^0)'\widehat{z}_T + N^{-1/2}\sqrt{N}(\overline{\lambda}^{-1\prime}\widehat{F}_T - F_T)$$

where $\delta^0 = [\alpha'\overline{\lambda}^{-1\prime}, \beta']'$.

- Problem: $\widehat{\delta}$ is not necessarily consistent for δ^0 when $\mathrm{rk}\,\overline{\lambda} = r \leq m$!

- Reason: When $r \leq m$ we can show that there is an $m \times m$ positive definite rotation matrix $\overline{\Lambda}$ such that

$$\overline{\Lambda}'\widehat{F}_t = \left[\begin{array}{c} F_t \\ 0_{(m-r)\times 1} \end{array}\right] + o_p(1)$$

Asymptotics

- This means that $T^{-1}\sum_{t=1}^{T-h}\widehat{z}_t\widehat{z}_t'$ – the "signal matrix" in $\widehat{\delta}$ – is asymptotically singular.

- In spite of this, we have

$$t(y_{T+h|T}) = \frac{\widehat{y}_{T+h|T} - y_{T+h|T}}{\sqrt{T^{-1}\phi^0 + N^{-1}\Phi^{0\prime}\Sigma_e\Phi^0}} \to_d N(0,1)$$

where $\Sigma_e = \lim_{N,T\to\infty} NE(\bar{e}_T\bar{e}_T')$, and ϕ^0 and Φ^0 are given in the paper.

- It follows that

$$\min\{\sqrt{N},\sqrt{T}\}(\widehat{y}_{T+h|T} - y_{T+h|T}) = O_p(1)$$

Asymptotics

- Inference requires estimators of ϕ^0 and $\Phi^{0\prime}\Sigma_e\Phi^0$.
- We propose using $\widehat{\phi}$ and $\widehat{\alpha}'\widehat{\Sigma}_e\widehat{\alpha}$, where $\widehat{\phi}$ and $\widehat{\Sigma}_e$ are given in the paper.
- We can show that $\widehat{\phi}$ and $\widehat{\alpha}'\widehat{\Sigma}_e\widehat{\alpha}$ are consistent if $r = m$.
- Hence, if $r = m$,

$$\begin{aligned}\widehat{t}(y_{T+h|T}) = \frac{\widehat{y}_{T+h|T} - y_{T+h|T}}{\sqrt{T^{-1}\widehat{\phi} + N^{-1}\widehat{\alpha}'\widehat{\Sigma}_e\widehat{\alpha}}} &= t(y_{T+h|T}) + o_p(1) \\ &\to_d N(0,1)\end{aligned}$$

Asymptotics

- Similarly, if we denote by $\mathrm{CI}_\gamma(y_{T+h|T})$ the $100 \cdot (1-\gamma)\%$ confidence interval for $y_{T+h|T}$, then

$$\lim_{N,T\to\infty} P(y_{T+h|T} \in \mathrm{CI}_\gamma(y_{T+h|T})) = \lim_{N,T\to\infty} P(|\widehat{t}(y_{T+h|T})| \leq z_{\gamma/2})$$
$$= 1-\gamma$$

- Problem: The inconsistency of $\widehat{\delta}$ causes $\widehat{\phi}$ and $\widehat{\alpha}'\widehat{\Sigma}_e\widehat{\alpha}$ to converge to random variables if $r < m$!

Asymptotics

- In spite of this, we can show that if $r \leq m$,

$$\lim_{N,T\to\infty} P(|\widehat{t}(y_{T+h|T})| > z_{\gamma/2}) \leq \gamma$$

- Hence, while $\widehat{t}(y_{T+h|T})$ is not asymptotically correctly sized, we know that it will not overreject!

- Confidence intervals will also be conservative;

$$\lim_{N,T\to\infty} P(y_{T+h|T} \in \mathrm{CI}_{\gamma}(y_{T+h|T})) \geq 1 - \gamma$$

Monte Carlo

- We set $h = 4$, $r = 1 < m = 2$, $\alpha = 1_{m\times 1}$, $W_1 = \cdots = W_T = 1$, $\beta = 1$, $\varepsilon_t \sim N(0,1)$ and $\lambda_i \sim (U[0,1], U[0,0.5])$.
- F_t is generated as

$$F_t = \rho F_{t-1} + \sqrt{1-\rho^2} u_t$$

  where $\rho = 0.5$ and $u_t \sim N(0,1)$.
- $e_{i,t} \sim N(0_{m\times 1}, \sigma_{e,i}^2 I_m)$, where $\sigma_{e,i}^2 \sim U[0.5, 1.5]$.

Monte Carlo

Table: Monte Carlo results for $\widehat{y}_{T+h|T}$.

		Coverage			MSE		
N	T	CA	PC	F	CA	PC	F
30	30	0.97	0.85	0.95	0.16	0.19	0.07
50	30	0.98	0.89	0.96	0.14	0.16	0.07
100	30	0.98	0.92	0.95	0.12	0.13	0.07
200	30	0.98	0.93	0.95	0.11	0.12	0.07
30	50	0.96	0.79	0.96	0.12	0.16	0.04
50	50	0.97	0.83	0.96	0.10	0.12	0.04
100	50	0.97	0.88	0.95	0.08	0.10	0.04
200	50	0.98	0.92	0.95	0.07	0.08	0.04
30	100	0.94	0.68	0.95	0.09	0.13	0.02
50	100	0.95	0.74	0.95	0.07	0.09	0.02
100	100	0.96	0.82	0.95	0.05	0.06	0.02
200	100	0.97	0.87	0.95	0.04	0.05	0.02
30	200	0.93	0.55	0.95	0.08	0.11	0.01
50	200	0.95	0.62	0.96	0.05	0.07	0.01
100	200	0.96	0.72	0.95	0.03	0.04	0.01
200	200	0.96	0.81	0.95	0.03	0.03	0.01

Empirical application

- We use the "usual" data set in the literature.
- We forecast the same eight macroeconomic variables as in Stock and Watson (JBES 2002).
- The panel data set can be divided into 14 categories.
- We take one average per category and use the BIC to select the ones to include in $\widehat{F}_t$.
- Predictors: $\widehat{z}_t = [\widehat{F}_t', 1, y_t]'$.

Empirical application

Table: MSE relative to AR $\times$ 100.

	$h = 6$		$h = 12$		$h = 24$	
Variable	CA	PC	CA	PC	CA	PC
IP	70.65	79.52	54.69	62.23	41.87	49.39
Income	70.04	76.59	60.21	62.09	60.34	66.55
Sales	74.80	84.70	58.30	63.72	39.86	43.93
Employees	75.99	83.78	52.13	58.35	37.42	39.03
CPI	67.35	68.96	66.44	74.83	65.50	88.48
Consumption	66.30	65.93	69.03	71.70	71.35	86.03
CPI less energy	71.98	68.79	73.25	82.87	76.81	99.12
Goods CPI	66.94	66.44	62.49	68.73	64.50	69.82

Conclusion

Thank you for listening!