Confidence intervals in regressions with estimated factors and idiosyncratic components

Jack Fosten

PII: S0165-1765(17)30220-3
DOI: http://dx.doi.org/10.1016/j.econlet.2017.05.034
Reference: ECOLET 7637

To appear in: *Economics Letters*

Received date : 16 February 2017
Revised date : 8 May 2017
Accepted date : 27 May 2017

- Recent models use both the factors and idiosyncratic components estimated from a big dataset
- We derive the distribution of the OLS estimates of these model parameters
- HAC standard errors must be adjusted to allow for the factor and idiosyncratic estimation error
- This is in contrast to existing results in the literature where estimation error vanishes when √T/N→0.

Confidence Intervals in Regressions with Estimated Factors and Idiosyncratic Components

Jack Fosten*
University of East Anglia, UK

May 8, 2017

Abstract

This paper shows that HAC standard errors must be adjusted when constructing confidence intervals in regressions involving both the factors and idiosyncratic components estimated from a big dataset. This result is in contrast to the seminal result of Bai and Ng (2006) where the assumption that $\sqrt{T}/N \to 0$ is sufficient to eliminate the effect of estimation error, where T and N are the time-series and cross-sectional dimensions. Simulations show vast improvements in the coverage rates of the adjusted confidence intervals over the unadjusted ones.

JEL Classification: C12, C22, C52, C53, C55

Keywords: Factor Model, Idiosyncratic Component, Inference, Confidence Intervals

[*] Department of Economics, University of East Anglia, Norwich, NR4 7TJ, UK. E-mail address: j.fosten@uea.ac.uk.

1 Introduction

In recent years, dynamic factor models have become a popular 'big data' method for applied econometricians wishing to make forecasts from large macroeconomic and financial datasets. Stock and Watson (2002) suggested to estimate a small number of factors from a large dataset, and use these factors in a second stage to augment standard forecasting models. The so-called "factor-augmented" forecasting model has subsequently achieved success in empirical studies; for a recent survey see Stock and Watson (2016). The validity of the two-stage factor-augmented model procedure was formally established by Bai and Ng (2006), who showed that the estimated factors can be treated as if they were the true, unobserved factors if $\sqrt{T}/N \to 0$, where T and N are the number of time series observations and variables respectively. This seminal result showed that confidence intervals for ordinary least squares (OLS) estimates of pure factor-augmented models can be constructed in the usual way by implementing heteroskedasticity and autocorrelation (HAC) robust standard errors. Only recently has this come under question in a sequence of papers (Gonçalves and Perron, 2014; Djogbenou et al., 2015; Gonçalves et al., 2017) which showed that factor estimation error *can* affect statistical inference in an asymptotic framework where $\sqrt{T}/N \to c$ where $c > 0$.

In this paper we show that the confidence intervals of Bai and Ng (2006) may be incorrect, even under the assumption that $\sqrt{T}/N \to 0$, when estimated idiosyncratic components from the factor model are additionally included in the second-stage forecasting regression. Models involving the estimated idiosyncratic components have been recently studied in papers such as Luciani (2013), Engel et al. (2015) and Fosten (2017) for forecasting macroeconomic and exchange rate series, but the asymptotic distribution of the regression estimates has not been derived for these models to the best of our knowledge.

The main finding of this paper is that an additional term appears in the variance-covariance matrix of the OLS coefficient estimates for the factors, relative to that of Bai and Ng (2006). This result is caused by the additional source of estimation error which arises from including the idiosyncratic components in the model. Without making an adjustment to the standard errors, they may be severely underestimated and the resulting confidence intervals will be too narrow, thereby invalidating statistical inference. Simulation results show very good coverage rates for confidence intervals constructed using the adjusted standard errors proposed in this paper, whereas coverage can be very poor when using the unadjusted confidence intervals of Bai and Ng (2006). This result is an important step for further study looking at topics such as bootstrap inference, or forecast intervals based on these models.

2 Set-up

The forecasting model for predicting y_{t+h} at the forecast horizon $h > 0$ using a set of N variables X_t is:

$$y_{t+h} = \beta^{0\prime} F_t^0 + \alpha^{0\prime} u_t^0 + \varepsilon_{t+h} \tag{1}$$

where we assume that X_t has the factor structure:

$$X_t = \Lambda F_t + u_t \tag{2}$$

In Equation (2), F_t is an $r \times 1$ vector of unknown factors, Λ is an $N \times r$ matrix of factor loadings (with typical row λ_i) and u_t is an $N \times 1$ vector of idiosyncratic error components. In the predictive regression model in Equation (1), $F_t^0 \subseteq F_t$ is an $r^0 \times 1$ subset of the factors and $u_t^0 \subseteq u_t$ is an $m^0 \times 1$ subset of the idiosyncratic components.[1] In other words, there are some subsets of both the factors and idiosyncratic components which affect y_{t+h}. Studies such as Boivin and Ng (2006) motivate the use of subsets of the factors by using only the 'real', 'nominal' or 'volatile' factors in predictive models. The use of subsets of the idiosyncratic components has been studied by Luciani (2013) and Fosten (2017).

The regression model in Equation (1) is not feasible as F_t^0 and u_t^0 are unknown in practice, but they can both be estimated from the factor model in Equation (2) after placing some identifying restrictions. Stock and Watson (2002) suggest to use standard Principal Components Analysis (PCA) which sets $\widehat{F}$ to be the r eigenvectors corresponding to the r largest eigenvalues of the $T \times T$ matrix XX'/TN. By normalising $\widehat{F}'\widehat{F}/T$, this gives rise to the estimate $\widehat{\Lambda} = X'\widehat{F}/T$ which in turn yields the estimate of the idiosyncratic component vector $\widehat{u}_t = X_t - \widehat{\Lambda}\widehat{F}_t$. The vectors $\widehat{F}_t^0$ and $\widehat{u}_t^0$ are then accordingly obtained as subsets of $\widehat{F}_t$ and $\widehat{u}_t$. Under standard PCA, the factor estimates are consistent for the true factors up to the rotation matrix $H = \widehat{V}^{-1}(\widehat{F}'F/T)(\Lambda'\Lambda/N)$ where $\widehat{V}$ is the $r \times r$ matrix of the r largest eigenvalues of XX'/TN. However, Bai and Ng (2013) propose alternative identifying assumptions which yield a rotation matrix which is asymptotically equal to a matrix with ±1 on the principal diagonal.

Letting $\widehat{\beta}^0$ and $\widehat{\alpha}^0$ be the OLS estimates from regressing y_{t+h} onto $\widehat{Z}_t^0 = [\widehat{F}_t^{0\prime}, \widehat{u}_t^{0\prime}]'$, and by denoting the $(r^0 + m^0) \times 1$ vectors $\widehat{\theta}^0 = [\widehat{\beta}^{0\prime}, \widehat{\alpha}^{0\prime}]'$ and $\theta^0 = [\beta^{0\prime}(H^0)^{-1}, \alpha^{0\prime}]'$, where H^0 is the relevant $r^0 \times r^0$ sub-matrix of the rotation matrix H,[2] we are interested in deriving the asymptotic distribution of:

$$\widehat{\theta}^0 = \left(\widehat{Z}^{0\prime}\widehat{Z}^0\right)^{-1} \widehat{Z}^{0\prime} y \tag{3}$$

where $\widehat{Z}^0$ is the $T \times (r^0 + m^0)$ equivalent of $\widehat{Z}_t^0$ in matrix form and similarly y is a $T \times 1$ vector. The distribution of $\widehat{\theta}^0$ has not yet been derived in the literature and requires dealing with the estimation error arising from both the estimated factors and idiosyncratic components.

3 Asymptotics

In deriving the asymptotic behaviour of $\widehat{\theta}^0$, there are several assumptions required to be placed onto the factors, loadings and idiosyncratic components. Since these assumptions are numerous

[1] We could also include a vector of 'must-have' regressors, W_t, such as a constant or lags of y_t, as in Bai and Ng (2006) and Gonçalves and Perron (2014), but these are omitted here for simplicity of notation.

[2] Using the identification schemes of Bai and Ng (2013), where the limit of H is an $r \times r$ matrix of ±1, the limit of H^0 will simply be the $r^0 \times r^0$ matrix of ±1 corresponding to the subset F^0 within F.

and similar in nature to the assumptions of Bai and Ng (2006) and Gonçalves and Perron (2014), for the sake of brevity these have been relegated to the Online Appendix to this paper. The assumptions essentially limit the moments and dependence of λ_i, F_t and u_{it}. We modify the Bai and Ng (2006) assumptions so that the identification PC1 of Bai and Ng (2013) holds (Assumption 2), which is helpful in the case where $r > 1$ in ensuring that the rotation matrix is asymptotically equal to a diagonal matrix with ±1 on the principal diagonal. This guarantees that subsets of estimated factors are consistent for the relevant subset of the true factors F_t^0, and not a linear combination of all of all r true factors F_t.[3] The only other modifications to the assumptions of Bai and Ng (2006) are in Assumption 5 where we extend theirs to allow the subset of the idiosyncratic components u_t^0 to enter the regression alongside the factors. We also allow for heteroskedasticity and autocorrelation in the variance-covariance matrices, as in Gonçalves and Perron (2014), which simply determines whether we require an estimator which is robust only to heteroskedasticity as in Bai and Ng (2006), or to both heteroskedasticity and autocorrelation as in Gonçalves and Perron (2014).

The following theorem outlines the asymptotic distribution for $\widehat{\theta}^0$:

Theorem 1. *Under Assumptions 1 to 5 detailed in the Online Appendix, and if* $\sqrt{T}/N \to 0$ *as* $N, T \to \infty$*, then:*

$$\sqrt{T}\left(\widehat{\theta}^0 - \theta^0\right) \xrightarrow{d} N\left(0, \Sigma_\theta\right)$$

where the variance covariance matrix Σ_θ *has the form:*

$$\Sigma_\theta = \left(\Phi^{0\dagger}\Sigma_{ZZ}\Phi^{0\prime\dagger}\right)^{-1}\left(\Phi^{0\dagger}\Sigma_{Z\epsilon}\Phi^{0\prime\dagger} + \begin{bmatrix} V_{F^0F}V_{Fu^0\alpha^0}V'_{F^0F} & \underset{r^0\times m^0}{0} \\ \underset{m^0\times r^0}{0} & \underset{r^0\times r^0}{0} \end{bmatrix}\right)\left(\Phi^{0\dagger}\Sigma_{ZZ}\Phi^{0\prime\dagger}\right)^{-1} \tag{4}$$

where $\Phi^{0\dagger} = p\lim\left(\Phi^0\right) = p\lim(diag(H^0, I_{m^0})) = diag(H^{0\dagger}, I_{m^0})$, $\Sigma_{Z\varepsilon} = \lim_{T\to\infty} Var\left[\frac{1}{\sqrt{T}}\sum_{t=1}^T Z_t^0\varepsilon_{t+h}\right]$, $\Psi(\alpha^0) = \lim_{T\to\infty} Var\left[\frac{1}{\sqrt{T}}\sum_{t=1}^T F_t u_t^{0\prime}\alpha^0\right]$, $V_{F^0F} = p\lim H^0\left(\frac{1}{T}\sum_{t=1}^T F_t^0 F_t'\right)H' = H^{0\dagger}\Sigma_{F^0F}H'^{\dagger}$ *and* $V_{Fu^0\alpha^0} = \lim Var\left(H\frac{1}{\sqrt{T}}\sum_{t=1}^T F_t u_t^{0\prime}\alpha^0\right) = H^\dagger\Psi(\alpha^0)H^{\dagger\prime}$*, where* $H^\dagger = p\lim H = diag(\pm 1)$*, and* $H^{0\dagger}$ *is the sub-matrix of* $H^\dagger$ *corresponding to* F_t^0.

A heteroskedasticity and autocorrelation (HAC) robust estimator of Σ_θ *is given by:*

$$\widehat{\Sigma}_\theta = \left(\frac{1}{T}\sum_{t=1}^T \widehat{Z}_t^0\widehat{Z}_t^{0\prime}\right)^{-1}\left(\widehat{\Sigma}_{Z\varepsilon} + \begin{bmatrix} \left(\frac{1}{T}\sum_{t=1}^T \widehat{F}_t^0\widehat{F}_t^{0\prime}\right)\widehat{V}_{Fu^0\alpha^0}\left(\frac{1}{T}\sum_{t=1}^T \widehat{F}_t^0\widehat{F}_t^{0\prime}\right)' & \underset{r^0\times m^0}{0} \\ \underset{m^0\times r^0}{0} & \underset{r^0\times r^0}{0} \end{bmatrix}\right)\left(\frac{1}{T}\sum_{t=1}^T \widehat{Z}_t^0\widehat{Z}_t^{0\prime}\right)^{-1}$$

where $\widehat{\Sigma}_{Z\varepsilon}$ *and* $\widehat{V}_{Fu^0\alpha^0}$ *are HAC estimators of* $\Phi^{0\dagger}\Sigma_{Z\epsilon}\Phi^{0\prime\dagger}$ *and* $V_{Fu^0\alpha^0}$*, which are respectively calculated using the long run variances of* $\frac{1}{\sqrt{T}}\sum_{t=1}^T \widehat{Z}_t^0\widehat{\varepsilon}_{t+h}$ *and* $\frac{1}{\sqrt{T}}\sum_{t=1}^T \widehat{F}_t\widehat{u}_t^{0\prime}\widehat{\alpha}^0$.

The proof of Theorem 1 can also be found in the Online Appendix. There is a crucial difference

[3] This assumption is also maintained by Gonçalves et al. (2015).

between the result of this theorem and that of Theorem 1 of Bai and Ng (2006), which is that in Equation (4) there is an additional term in the elements of the variance-covariance matrix which correspond to the estimated coefficient vector $\widehat{\beta}^0$ on the factors. This is apparent from inspection of the middle term in the equation for Σ_θ, which does not appear in the equivalent expression in Equation 3 of Bai and Ng (2006). The result is particularly interesting as it true even when $\sqrt{T}/N \to 0$, thereby overturning a key finding in the factor model literature where inference can proceed without adjusting the standard errors in the pure factor case with no idiosyncratic components.

The reason for this result is that the product of the factors with the estimation error term resulting from including the estimated idiosyncratic components produces a term which converges in distribution. This product appears in the OLS estimator of β^0 and therefore yields an additional term in its variance-covariance matrix which is not present in the pure factor-augmented model case in previous studies. On the other hand, no additional term appears in the variance of $\widehat{\alpha}^0$, the vector of coefficients on the idiosyncratic components. This is because the extra estimation error term converges in probability to zero when multiplied by the zero-mean idiosyncratic component, and so this part of the expression for the OLS estimator of α^0 does not contribute to its variance. A related result was found by Fosten (2017), although they did not derive the distribution of this term but noted that the convergence rate was different to the estimation error terms present in the pure factor case. The implication of Theorem 1 is that, if we do not adjust the standard errors on the factors accordingly, incorrect inference may be made and any resulting prediction intervals would correspondingly be incorrect.

4 Monte Carlo

We use a simple Monte Carlo set-up to demonstrate the finite sample properties of coverage rates for β^0, which must be calculated using adjusted standard errors as indicated by Theorem 1. We compare these coverage rates to those obtained by not adjusting the standard errors, which are only valid in the pure factor-augmented context of Bai and Ng (2006), and not in those involving idiosyncratic components. We take the simplest case where there is only one factor in X_t, in other words $r = 1$, and where the forecasting model for y_t contains the single factor and one idiosyncratic component, $m^0 = 1$, corresponding to the first idiosyncratic error u_{1t}:[4]

$$X_t = \Lambda F_t + u_t \tag{5}$$

$$y_{t+h} = 1 + F_t + u_{1t} + \varepsilon_{t+h} \tag{6}$$

The single factor is drawn as $F_t \sim iidN(0, 1)$, the loadings are an $N \times 1$ vector $\Lambda \sim iidN(1, 1)$, the idiosyncratic errors are the $N \times 1$ vector $u_t \sim iidN(0, K_F)$ and finally the forecast model errors are

[4] For this paper we assume the identity of u_{1t} to be known *a priori* but this can be consistently selected using the information criteria of Fosten (2017).

$\varepsilon_{t+h} \sim iidN(0, K_y)$.[5] Note that the use of a normal distribution for the loadings, rather than the uniform distribution, all but eliminates the issue of bias in the OLS estimators in the case where N is small, as discussed in Gonçalves and Perron (2014) and Ludvigson and Ng (2011). The common component ΛF_t therefore has unit variance, and so the signal-to-noise (STN) ratio in the factor model in Equation (5) is equal to $1/K_F$. Similarly in the forecast model in Equation (6), the STN ratio is $1/K_y$. These ratios will be varied in the simulation results.

Table 1: Simulation Results: Empirical Coverage Rates

Conf. Interval β^0 Unadjusted					Conf. Interval β^0 Adjusted				
1) Equal STN					1) Equal STN				
N/T	50	100	200	400	N/T	50	100	200	400
50	82.8%	85.5%	83.6%	83.6%	50	90.9%	92.6%	92.0%	91.9%
100	82.9%	84.0%	85.8%	84.0%	100	90.4%	90.8%	92.7%	92.4%
200	82.5%	82.9%	83.9%	85.0%	200	88.8%	91.7%	91.9%	92.0%
2) Low Factor Model STN					2) Low Factor Model STN				
N/T	50	100	200	400	N/T	50	100	200	400
50	78.2%	78.6%	78.2%	76.7%	50	90.9%	90.7%	91.7%	91.3%
100	78.3%	79.0%	80.9%	78.5%	100	91.0%	91.9%	93.4%	92.7%
200	78.6%	79.4%	80.2%	80.0%	200	91.7%	92.6%	92.3%	93.9%
3) Low Forecast Model STN					3) Low Forecast Model STN				
N/T	50	100	200	400	N/T	50	100	200	400
50	88.5%	89.5%	91.6%	89.6%	50	93.3%	93.5%	94.6%	93.3%
100	87.6%	87.1%	88.5%	90.5%	100	92.2%	91.4%	91.9%	94.9%
200	87.0%	87.3%	88.2%	90.2%	200	90.8%	92.1%	92.3%	93.7%

Notes: This table presents the empirical coverage rates for confidence intervals for β^0 using the adjusted and unadjusted standard errors. The nominal coverage rate is 95%.

We will explore the finite sample properties for a variety of sample sizes for the number of observations, T, and variables, N. Specifically we let $N = \{50, 100, 200\}$ and $T = \{50, 100, 200, 400\}$.[6] In order to see how the results are affected by different STN ratios in the factor model or forecast model, we run three scenarios. The first scenario has $K_F = K_y = 1$ to ensure an equal STN ratio in both Equation (5) and (6). The second scenario has $K_F = 2$ and $K_y = 1$ so that the factor model has a low STN ratio. The final scenario conversely has $K_y = 2$ and $K_F = 1$ so that the forecast model has a low STN ratio. We make $M = 1000$ Monte Carlo draws. Table 1 displays the results[7] for the coverage rates for β^0:

[5] Note that, as in the Monte Carlo study of Gonçalves and Perron (2014), this DGP with $r = 1$ means that assumption PC1 of Bai and Ng (2013) holds and so the factor is consistent up to the rotation ±1.

[6] We also ran some additional results with smaller sample sizes $T = 20$ and $N = 10, 20$. These results are not displayed here as they are somewhat smaller than existing simulations in the literature, but are available from the author on request.

[7] The results for α^0 are not reported as these confidence intervals do not require adjustments. Results are available on request.

The results clearly demonstrate the importance of adjusting the standard errors along the lines of Theorem 1. The results on the left of Table 1 show that the empirical coverage rates are very poor when using the unadjusted standard errors. The coverage rates are furthest from the nominal 95% rate in scenario 2 where the signal-to-noise ratio in the factor model is low, where empirical coverage is as low as 77%. There is generally improvement in the coverage rate as both N and T increase, as expected.[8] Turning to the results using the adjusted standard errors, we see that the coverage rates improve substantially, moving much closer to the 95% nominal coverage rate than the results with unadjusted standard errors.

5 Conclusion

This paper shows that caution must be taken when interpreting the coefficients from predictive models involving both the factors and idiosyncratic components estimated from a big dataset. We derive the asymptotic distribution of the OLS regression estimates, showing that it is not possible to apply standard HAC estimators of the standard errors in these models, as was suggested in the pure factor context of Bai and Ng (2006). This result is due to an additional term in the variance-covariance matrix of the OLS estimates for the factors, which is caused by the extra source of estimation error from the idiosyncratic component; something which holds even when $\sqrt{T}/N \to 0$. The result has important implications for further work, for example in forming prediction intervals based on these parameter estimates, or bootstrap methods for inference in these models, both of which are areas left for further study.

References

Bai, J. and S. Ng (2006). Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica 74*(4), 1133–1150.

Bai, J. and S. Ng (2013). Principal Components Estimation and Identification of Static Factors. *Journal of Econometrics 176*(1), 18–29.

Boivin, J. and S. Ng (2006). Are More Data Always Better for Factor Analysis? *Journal of Econometrics 132*(1), 169–194.

Djogbenou, A., S. Gonçalves, and B. Perron (2015). Bootstrap inference in regressions with estimated factors and serial correlation. *Journal of Time Series Analysis 36*(3), 481–502.

Engel, C., K. West, and N. Mark (2015). Factor Model Forecasts of Exchange Rates. *Econometric Reviews 34*(1), 32–55.

[8] For some configurations it is not the case that the coverage rates improve as N increases for a fixed T, however a similar finding can also be seen in the related simulations of Bai and Ng (2006) in their Table 1.

Fosten, J. (2017). Model Selection with Estimated Factors and Idiosyncratic Components. *Journal of Applied Econometrics (Forthcoming)*.

Gonçalves, S., M. W. McCracken, and B. Perron (2015). Tests of Equal Accuracy for Nested Models with Estimated Factors. *Federal Reserve Bank of St. Louis Working Paper 2015-025A*.

Gonçalves, S. and B. Perron (2014). Bootstrapping Factor-Augmented Regression Models. *Journal of Econometrics 182*(1), 156–173.

Gonçalves, S., B. Perron, and A. Djogbenou (2017). Bootstrap Prediction Intervals for Factor Models. *Journal of Business & Economic Statistics 35*(1), 53–69.

Luciani, M. (2013). Forecasting with Approximate Dynamic Factor Models: The Role of Non-pervasive Shocks. *International Journal of Forecasting 30*(1), 20–29.

Ludvigson, S. C. and S. Ng (2011). A Factor Analysis of Bond Risk Premia. In A. Ullah and D. Giles (Eds.), *Handbook of Empirical Economics and Finance*, pp. 313–372. Chapman and Hall.

Stock, J. H. and M. W. Watson (2002). Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association 97*(460), 1167–1179.

Stock, J. H. and M. W. Watson (2016). Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. In J. B. Taylor and H. Uhlig (Eds.), *Handbook of Macroeconomics*, Volume 2, pp. 415–525. Elsevier.