Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: http://www.tandfonline.com/loi/ubes20

Comment

Xu Cheng & Bruce E. Hansen

To cite this article: Xu Cheng & Bruce E. Hansen (2016) Comment, Journal of Business & Economic Statistics, 34:3, 345-347, DOI: 10.1080/07350015.2016.1189338

To link to this article: http://dx.doi.org/10.1080/07350015.2016.1189338

Published online: 19 Jul 2016.

Comment

Xu CHENG
Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104
(*xucheng@upenn.edu*)

Bruce E. HANSEN
Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706
(*bruce.hansen@wisc.edu*)

Factor-augmented regression is widely applied to forecasting in a data-rich environment. For most papers in the literature, a small number of factors are assumed to generate the comovement of a large number of predictors, see Stock and Watson (2002). In this inspiring article, Carrasco and Rossi look at the problem from a fresh angle: they relax this key assumption and provide a comprehensive investigation of various regularization methods under misspecified factor models. Their theoretical results derived under different types of sparsity conditions deepen our understanding of estimation and forecasting with large dataset and serve as important practical guidance.

This discussion follows the lead by Carrasco and Rossi and extends the study to several model averaging methods. First, we extend the simulation study in Cheng and Hansen (2015, hereafter CH) on forecast model selection and forecast model averaging by allowing, as in Carrasco and Rossi, for a large number of factors. We find that forecast combination by multi-step cross-validation (CV) is reasonably robust to the large factor setting. Second, we extend the simulation study in Carrasco and Rossi to include our proposed frequentist model averaging methods. In their simulation design, we find that model averaging (forecast combination) across factor models has lower MSFE than model selection in small samples, but the methods have similar MSFE in large samples.

We first summarize the frequentist model averaging methods considered by CH and second present the simulation results. Finally, we conclude with some additional remarks.

Frequentist Model Averaging. Following CH, we consider forecasting with the factor-augmented regression model $y_{t+h} = \alpha_0 + \alpha(L)y_t + \beta(L)'F_t + \varepsilon_{t+h}$, where $h \geq 1$ is the forecast horizon, $\alpha(L)$ and $\beta(L)$ are lag polynomials of order p and q, respectively, and $F_t \in \mathbb{R}^r$ are unobserved common factors satisfying $X_{it} = \lambda_i' F_t + e_{it}$. As discussed by Bai and Ng (2009) and Carrasco and Rossi, important factors in the panel may not be important for forecasting. CH treats the structures of $\alpha(L)$ and $\beta(L)$ as unknown and introduce methods to select the factors and lag structures for forecasting. Assuming $p \leq p_{\max}$ and $q \leq q_{\max}$ for some large numbers $p_{\max}$ and $q_{\max}$, the largest possible approximating model includes the regressors $z_t = (1, y_t, \ldots, y_{t-p_{\max}}, F_t', \ldots, F_{t-q_{\max}}')'$. Now suppose that the forecaster is considering M approximating models indexed by $m = 1, \ldots, M$, where each approximating model m specifies a subset $z_t(m)$ of the regressors z_t. The forecaster's mth approximating model is then $y_{t+h} = z_t(m)'b(m) + \varepsilon_{t+h}(m)$.

For estimation, we replace the unobservable factor F_t by its principle component estimate $\widehat{F}_t$. Let $\widehat{b}(m)$ denote the least-squares estimate of $b(m)$ based on the mth approximating model and $\widehat{\varepsilon}_{t+h}(m) = y_{t+h} - \widetilde{z}_t(m)'\widehat{b}(m)$ denote the residual. The least-squares forecast of y_{T+h} by the mth approximating model is $\widehat{y}_{T+h|T}(m) = \widetilde{z}_T(m)'\widehat{b}(m)$. Forecast combinations take the form $\widehat{y}_{T+h|T}(w) = \sum_{m=1}^M w(m)\widehat{y}_{T+h|T}(m)$, where $0 \leq w(m) \leq 1$ for $m = 1, \ldots, M$ and $\sum_{m=1}^M w(m) = 1$, or equivalently that $w = (w(1), \ldots, w(M))' \in \mathcal{H}^M$, the unit simplex in $\mathbb{R}^M$. We choose w by the Mallows averaging criterion (Hansen 2007) and leave-h-out CV criterion (Hansen 2010; Hansen and Racine 2012).

The Mallows averaging criterion is

$$C_T(w) = \frac{1}{T}\sum_{t+h=1}^{T}\left(\sum_{m=1}^{M} w(m)\widehat{\varepsilon}_{t+h}(m)\right)^2 + \frac{2\widehat{\sigma}_T^2}{T}\sum_{m=1}^{M} w(m)k(m), \quad (1)$$

where $k(m)$ denotes the number of regressors in the mth model and $\widehat{\sigma}_T^2$ is a preliminary estimate of $\sigma^2 = \mathbb{E}\varepsilon_t^2$. The Mallows selected weight vector is $\widehat{w} = \operatorname{argmin}_{w\in\mathcal{H}^M} C_T(w)$.

To apply leave-h-out CV Averaging, let $\widetilde{\varepsilon}_{t,h}(m)$ be the residual obtained by least-squares estimation of the m^{th} model with $\{t-h+1, \ldots, t+h-1\}$ omitted. For forecast combination, the leave-h-out prediction residual is $\widetilde{\varepsilon}_{t+h,h}(w) = \sum_{m=1}^M w(m)\widetilde{\varepsilon}_{t+h,h}(m)$ and the leave-h-out CV criterion is

$$CV_{h,T}(w) = \frac{1}{T}\sum_{t=1}^{T}\widetilde{\varepsilon}_{t+h,h}(w)^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{m=1}^{M} w(m)\widetilde{\varepsilon}_{t+h,h}(m)\right)^2. \quad (2)$$

The CV selected weight vector is $\widehat{w} = \operatorname{argmin}_{w\in\mathcal{H}^M} CV_{h,T}(w)$. For both Mallows averaging and leave-h-out CV averaging, replacing $\mathcal{H}^M$ by $w(m) \in \{0, 1\}$ and $\sum_{m=1}^M w(m) = 1$ changes the model averaging method to the corresponding model selection method.

Simulation Design in Cheng and Hansen (2015). We first consider the simulation design in CH, which follows a

Journal of Business & Economic Statistics
July 2016, Vol. 34, No. 3
DOI: 10.1080/07350015.2016.1189338
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

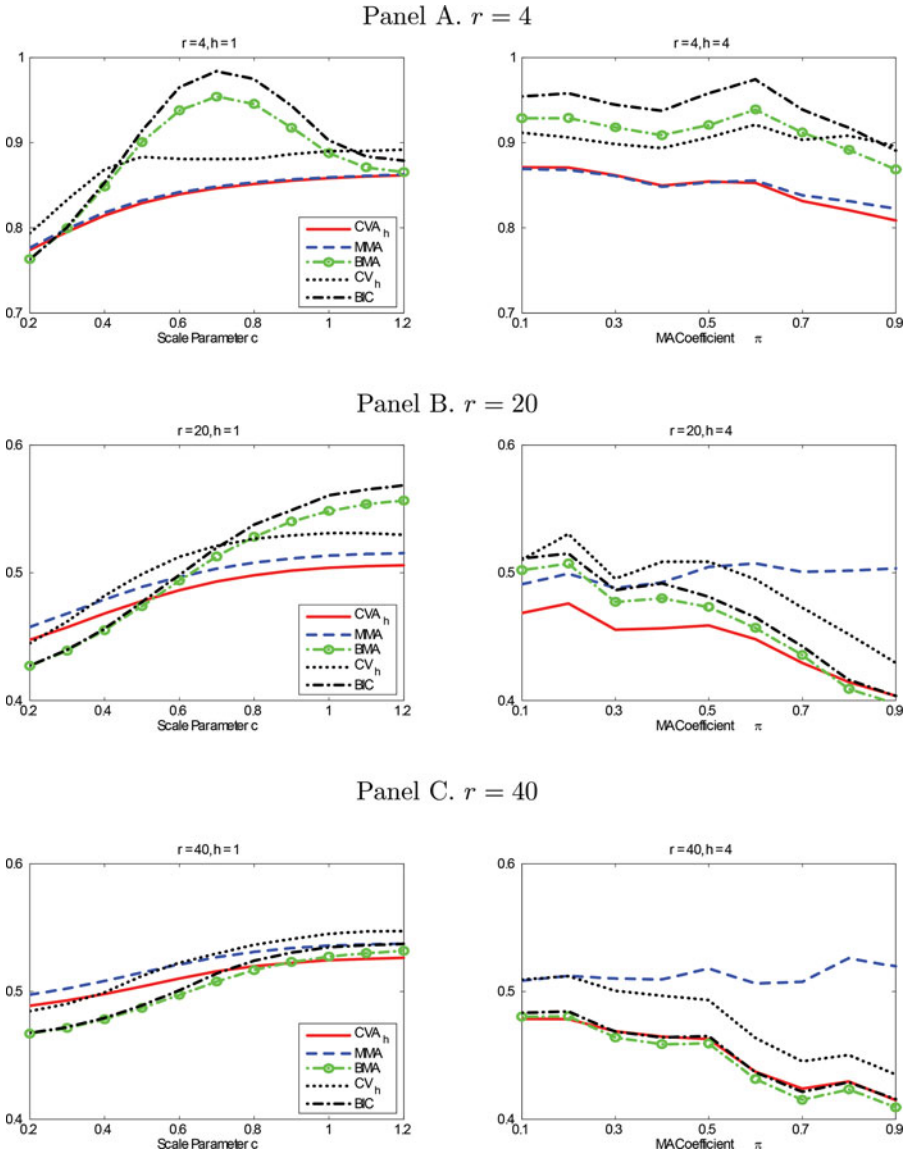

Figure 1. MSFE for simulation design in Cheng and Hansen (2015). Note: The MSFE of the least-squares forecast with all regressors is normalized to be 1. CVA_h is leave-h-out cross-validation averaging. MMA is Mallows model averaging. BMA is Bayesian model averaging. CV_h is model selection with leave-h-out cross-validation. BIC is model selection with Bayesian information criterion.

similar design in Bai and Ng (2009). The data-generating process (DGP) of the panel data is $x_{it} = \lambda_i' F_t + \sqrt{r} e_{it}$, the jth component of F_t follows from $F_{jt} = \alpha_j F_{jt-1} + u_{jt}$, $e_{it} = \rho_i e_{it-1} + \xi_{it}$, and the DGP for the forecast regression equation is $y_{t+h} = \beta'(F_{2t}, F_{4t}, F_{2t-1}, F_{4t-1}, F_{2t-2}, F_{4t-2}) + \varepsilon_{t+h}$ and $\varepsilon_{t+h} = v_{t+h} + \pi v_{t+h-1} + \cdots + \pi^{h-1} v_{t+1}$, where $(u_{jt}, \xi_{it}, v_{t+h}, \ldots, v_{t+1}) \sim N(0, I_{h+2})$ is iid over t, for all j and i. We consider $r = 4, 20, 40$ and $h = 1, 4$ in the cases presented in Figure 1. The parameter values are $\alpha_j \sim U[0.2, 0.8]$, $\rho_i \sim U[0.3, 0.8], \lambda_i \sim N(0, rI_r)$, $\beta = c \cdot (0.5, 0.5, 0.2, 0.2, 0.1, 0.1)$. The constant c is a scaling parameter ranging from 0.2 to 1.2 for $h = 1$. For multi-step-ahead forecasting, the moving average parameter π ranges from 0.1 to 0.9 and the scale parameter c is held at 1. The sample size is $N = 100$ and $T = 100$. For lag length choices, $p\max = 4$ and $q\max = 4$. The number of simulation repetitions is 10,000. The true number of factors is unknown in practice, therefore we start by selecting the number of factors using the information criterion IC_{p2} recommended by Bai and Ng (2002). The mean squared forecast error (MSFE) of least-squares forecast with all regressors is normalized to be 1.

Panel A of Figure 1 is the standard case where $r = 4$, much smaller than $N = T = 100$. For $h = 1$, CH show that Mallows model averaging and leave-h-out CV averaging perform better than Bayesian model averaging (BMA)[1], simple averaging with equal weights, leave-h-out CV model selection, Mallows model selection, AIC, and BIC.[2] For $h = 4$, CH show that leave-h-out CV averaging performs best when the forecast error exhibits strong serial dependence. These patterns are demonstrated in Panel A. In an empirical study, CH further show that leave-h-out

[1]Our BMA weights are set as $w(m) = \exp(-\text{BIC}(m)/2)/\sum_{i=1}^{M} \exp(-\text{BIC}(i)/2)$, where BIC($m$) is the BIC for the mth model. This is an approximate BMA for the case of equal model priors, and diffuse model priors on parameters.

[2]Some results are not plotted due to space limit.

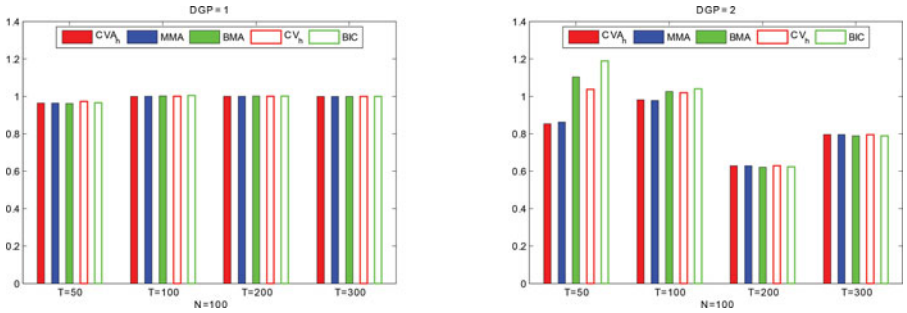

Figure 2. MSFE for Simulation Design in Carrasco and Rossi (2016).

CV averaging compares favorably to many shrinkage methods considered by Stock and Watson (2012).

Panels B and C of Figure 1 consider $r = 20$ and $r = 40$, respectively. These results extend the study in CH to misspecified factor models because the theory in CH assumes that r is much smaller than N and T. For $r = 20$ and $h = 1$ we find mixed results, with the MSFE of BMA and BIC sensitive to the signal-to-noise ratio (c). They have lower MSFE for small c but leave-h-out CV averaging has lower MSFE for larger c. For $h = 4$ the CV averaging method has the lowest MSFE for nearly all signal-to-noise values. For $r = 40$, BMA and BIC have improved performance relative to leave-h-out CV averaging, though the methods are nearly equivalent for $h = 4$.

Simulation Design in Carrasco and Rossi (2016). Here, we consider DGP1 and DGP2 in Carrasco and Rossi to compare different procedures under $r = 4$ and $r = 50$. We consider $N = 100$ and $T = 50, 100, 200, 300$ for $h = 1$. The MSFE is computed as the average over 10,000 simulation repetitions, where for each sample periods 1 to $T - 1$ are used to forecast period T. No lags of Y_t or F_t are used in this forecasting. The largest number of factors $r$max follows that in Carrasco and Rossi.

Figure 2 plots the MSFE relative to the least-squares forecast with all regressors, where the number of factors is estimated by the Bai and Ng (2002) criterion PC_{p2} as in Carrasco and Rossi. For DGP1 where $r = 4$, all methods work equally well as expected. For DGP2 with 50 relevant factors and $T = 50$, we find that our recommended leave-h-out CV has meaningfully lower MSFE than BMA and BIC selection. These advantages diminish for $T = 100$ and disappear for $T = 200$ and 300, where all methods have similar MSFE.[3] Table 2 in Carrasco and Rossi suggests that some of their procedures, in particular the Landweber Fridman method, have even better forecasting performance than the methods considered in Figure 2 for $r = 50, N = 100$ and $T = 50$.

The message from Figure 2 is that once again there are large potential gains from model averaging across factor models, even in the high-dimensional context of large factor models. We find the most substantial reductions in MSFE for the case of small samples ($T = 50$), but this is precisely the context where the gains are most important.

Conclusion and Discussion. This thought-provoking article by Carrasco and Rossi has shown that it is important to consider methods that are robust to meaningful perturbations from classical assumptions on models for large-scale data. Our simulation studies show that the CV averaging method continue to dominate various popular alternatives in models with a large number of factors, robustifying the observation in CH in the direction pointed by Carrasco and Rossi. On the other hand, some regularization methods coupled with the tuning parameter choice proposed by Carrasco and Rossi have shown even greater potential when the sample size is moderate. Combining these two observations, it would be interesting to see in future research (i) whether the leave-h-out CV criterion could be used for tuning parameter selection in the context of the proposed methods of Carrasco and Rossi; and (ii) whether further improvements could be obtained by forecast combination across classes of regularization methods, in particular by using the leave-h-out CV criterion for weight selection.

[3]We also have considered using Bai and Ng (2002) IC_{p2} criterion to select the number of factors, as in CH. For DGP 1 where $r = 4$, these two criteria have similar performances. For DGP 2 where $r = 50$, and thus the factor model is severely misspecified, IC_{p2} produces a much smaller estimate for $T = 50$ and $N = 100$. Sometimes IC_{p2} selects 0 factor in this misspecified case. When this happens, we apply all procedures in Figure 2 with the number of factors set to 10. Overall, we see the same pattern as that in Figure 2: Mallows model averaging and CV averaging have some advantages over the other procedures for $T = 50$ and $T = 100$ and these advantages disappear for larger T.

REFERENCES

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [346,347]

——— (2009), "Boosting Diffusion Indices," *Journal of Applied Econometrics*, 4, 607–629. [345]

Carrasco, M., and Rossi, B. (2016), "In-Sample Inference and Forecasting in Misspecified Factor Models," *Journal of Business & Economic Statistics*, 34, 313–338. [347]

Cheng, X., and Hansen, B. E. (2015), "Forecasting With Factor-Augmented Regression: A Frequentist Model Averaging Approach," *Journal of Econometrics*, 186, 280–293. [345]

Hansen, B. E. (2007), "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189. [345]

——— (2010), *Multi-Step Forecast Model Selection*, University of Wisconsin, Working paper. [345]

Hansen, B. E., and Racine, J. S. (2012), "Jackknife Model Averaging," *Journal of Econometrics*, 167, 38–46. [345]

Stock, J. H., and Watson, M. W. (2002), "Forecasting Using Principle Components From a Large Number of Predictors," *Journal of American Statistical Association*, 97, 1167–1179. [345]

——— (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business and Economic Statistics*, 30, 481–493. [347]