Online Appendix for ``The Impact of Big Data on Firm Performance: an Empirical Investigation" by Patrick Bajari, Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki

1. Impact of Data Size on Quality of Forecasting in a Theoretical Model

Retail companies use proprietary forecasting models, which are often the result of a complicated engineering effort, tailored to the particularities of the business. Theoretical properties of such complicated forecasting models may never be understood completely. Hence we give a theoretical benchmark based on a well-known, state-of-the-art model used in academia to model panel time series and forecasting. In particular, we utilize the Augmented Factor model; see, for example, Bai and Ng (2002), Bernanke et al (2004), and Bai (2009). Within this fairly general, yet tractable model, we ask the question of how forecast errors would be determined and what role do the data size, the number of products and number of time periods in the available history, have.

The quantity $Q_{i,t}$ sold by a retailer of a product i at time t, obeys the following equation:

$$\log\big((Q_{i,t}+1)/V_{i,t}\big) = Q^0_{i,t} = \alpha'_i F_t + X'_{i,t}\beta + \epsilon_{i,t},\ i = 1,\dots,N,\quad t = 1,\dots,T,$$

where $V_{i,t} \geq 1$ is a known velocity variable, which describes the stochastic, possibly non-stationary level of the series, and $Q^0_{i,t} \geq 1$ is a reference demand level, where $\{Q^0_{i,t}\}_{t=0}^{\infty}$ stationary for each i. Note that $Q_{i,t} = 0$ if and only if $V_{i,t} = 1$ and $Q^0_{i,t} = 1$. The quantity $Q_{i,t}$ (plus 1) is determined by the base demand times the multiplier $V_{i,t}$ that captures the ``size of the firm" in product i sales as well as "velocity" of the product. Velocity $V_{i,t}$ reflects the notional popularity of the product i at time t, and represents the product-specific size of the retailer, as specific to the product. A simple example of $V_{i,t}$ is given by the lagged sales $V_{i,t} = (Q_{i,t-1} + 1)$. We can normalize the velocity at $V_{i,t}$ =1 for $t = 1$. The vector of velocities $\{V_{i,t}\}_{i=1}^{N}$ characterizes the overall ``size of the retailer" over time. In the model, the base quantity index $Q^0_{i,t}$ is determined by latent factor components plus observed components plus stochastic shocks: (i) Time-varying factors F_t are latent time-varying common factors, such as macro-economic factors, seasonality, and fashion factors. (ii) The product varying factors α_i are the product-specific loadings on the latent factors F_t, to which product demand responds differently. They include the conventional fixed effects model, when $F_t = 1$. (iii) The observed component is determined by a p-dimensional vector $X_{i,t}$ of observed time-varying product features, such as prices of the product

as well as its substitutes and complements, multiplied by a common parameter β. (iv) The latent shocks $\epsilon_{i,t}$ are unobserved, unlearnable error components.

We define the quality of forecast in relative terms,

$$relative\ erro_{\ i,t} \coloneqq \frac{\left|\left(\hat{Q}_{i,t}+1\right)-\left(Q_{i,t}+1\right)\right|}{\left(Q_{i,t}+1\right)}$$

where $\hat{Q}_{i,t}$ is the forecast constructed based on the estimators as in Bai and Ng (2002) and Bai (2009); See Bajari et al. (2018) for details. Then it follows that under suitable regularity conditions $relative\ error_{i,t}$ is of stochastic order $\sqrt{1/N}+\sqrt{1/T}$. Under regularity conditions, that implies that

$$\Pr\left(\left|relative\ error_{i,t}\right|>c\right)\leq \Lambda + \mathrm{C}(\sqrt{1/N}+\sqrt{1/T}),$$

for some constants c, C and Λ that do not depend on the data size (N,t). This bound motivates our empirical specification,

2. Estimation Results of Fixed Effects Model

	Dependent variable:		
	1(Relative Forecast Error > X)		
	Without Trend	Time Trend	Time Effects
	(1)	(2)	(3)
Age > 20	-0.800***	-0.029	-0.140
	(0.046)	(0.086)	(0.085)
(Age>20) Inv Root Age	9.230***	0.395	1.460
	(0.574)	(0.948)	(0.936)
(Age > 20) Inverse Age	-27.472***	-1.717	-4.247
	(1.826)	(2.678)	(2.643)
N > 200	-0.487**	0.053	0.292
	(0.212)	(0.229)	(0.237)
(N > 200) Inverse Root N	8.534*	-6.234	-13.931**
	(4.608)	(5.106)	(5.422)
(N > 200) Inverse N	-23.457	81.345	153.613**
	(62.620)	(65.474)	(67.159)
Trend		-0.327***	
		(0.075)	
Squared Trend		-0.060	
		(0.045)	
Observations	496,259	496,259	496,259
R^2	0.276	0.278	0.284
Adjusted R^2	0.268	0.270	0.277

Note: Standard errors are clustered by product and date.

References

Bai, Jushan. ``Panel Data Models with Interactive Fixed Effects." *Econometrica*, 2009, 77(4), 1229-352.

Bai, Jushan, and Serena Ng. "Determining the number of factors in approximate factor models." *Econometrica*, 2002, 70(1), 191-221.

Bajari, Patrick L., Victor Chernozhukov, Ali Hortaçsu, and Junichi Suzuki. ``The Impact of Big Data on Firm Performance: an Empirical Investigation.” 2018, NBER working paper no. 24334.

Bernanke, Ben S., Jean Boivin, and Piotr Eliasz, ``Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach." January 2004, NBER working paper no. 10220.