Comment

Norman R. Swanson
Department of EconomicsRutgers University, New Brunswick, NJ 08901 (*nswanson@econ.rutgers.edu*)

1. DATA MINING AND FACTOR ANALYSIS: A BRIEF HISTORICAL PERSPECTIVE

Data mining in economics has become an important topic. A key reason for this is that so-called "big data" dominates the data-rich environments in which empirical economists operate. This holds true in fields ranging from macroeconomics, where hundreds of time series are often jointly modeled, to finance, where interest lies in analyzing high frequency data. Of course, big data environments have long been available in various other areas in economics. For example, there are many data-rich surveys available to practitioners, such as the PSID, which is a large longitudinal panel study begun in 1968. With so much data now available, attention has turned in recent years to the refinement of old (but relevant) empirical tools and the development of new ones. One of the oldest empirical tools is factor analysis. As is widely known, in 1904 Charles Spearman posited that there was a hidden or latent structure underlying human intelligence. In his model, he assumed that there was one latent common factor, called general intelligence (say G_i). He argued that particular mental abilities (features) for person i, say $X_{i,j}$ could be predicted using the model

$$X_{i,j} = \beta_j G_i + \varepsilon_{i,j}.$$

Assuming that the common factor is zero mean and that $\text{corr}(\varepsilon_{i,j} G_i) = 0$, for any j implies that the correlation between feature $X_{\cdot,j}$ and G is simply β_j. Thus, features $X_{\cdot,j}$ and $X_{\cdot,k}$ have correlation equal to the product of their factor loadings (say $\rho_{jk} = \beta_j \beta_k$). This in turn implies that for any four features, j, k, l, and m

$$\rho_{jl}\rho_{km} = \rho_{jm}\rho_{kl}.$$

Interestingly, Spearman (1904) found that this so-called tetrad equation was a reasonable approximation when analyzing his school grades dataset; and hence he argued that his single factor model was correctly specified. Although later work showed that this equation does not hold, the seeds were sown, and factor analysis has thrived ever since. Moreover, the tetrad equation became very important in early work on causality and causal graphs, such as Wold causal chain (see, e.g., Wright 1921; Wold 1954; Simon 1955), and some decades later in early artificial intelligence research (see, e.g., Glymour et al. 1987; Pearl and Verma 1994; Swanson and Granger 1997; Pearl 2000). This work in turn underpins some of the most recent work on machine learning in numerous disciplines ranging from neuroscience to engineering. In summary, the interesting developments in factor analysis, machine learning, variable selection, and shrinkage discussed in CR have direct roots stretching back over 100 years.

2. CONTRIBUTIONS OF CARRASCO AND ROSSI (2016)

This well-crafted article delivers on a number of fronts, both theoretical and methodological. In this discussion, I will attempt to summarize the contributions of three main parts of the article.

In the first part of the article, the authors derive asymptotic properties and carry out Monte Carlo and empirical investigations of four data reduction (i.e., data mining, variable selection, and shrinkage) methods, including principal components analysis (PCA), ridge regression (RR), Landweber-Fridman data reduction (LR), and partial least squares (PLS). In this context, they examine two models—an "ill-posed" model where the eigenvalues of $X'X/T \to 0$, as $N \to \infty$ (see below for definition of X), and a factor model with the number of latent factors, r, fixed. One key finding is that the convergence rates of the prediction errors associated with the four methods differ by method and model. For example, in the case of the ill-posed model, RR, LF, and PCA are characterized by the same rate in many cases. However, in case of the factor model, PCA and LF rates are faster than RR. Also, there is a bias and variance trade-off when calculating the rate for PLS, regardless of model. (The authors also compare these methods with a spectral cut-off type PCA, which is called SC in the article. When constructing factors, SC uses a thresholding method to select eigenvectors (see sec. 2 of their article for further details).) These theoretical findings are largely supported via extensive Monte Carlo and empirical investigations.

In the second part of their article, Carrasco and Rossi focus on the tuning or regularization parameter that characterizes the estimation methods that they examine. For example, they note that (in forecasting contexts) the key tuning parameter in PCA is the number, k, of principal components. They propose using generalized cross-validation (GCV) and related statistics to choose k (see below for further discussion), and argue that their approach is sensible, and is a good alternative to extant methods including the Bai and Ng (2002) statistics, because GCV, minimizes forecast prediction error associated with a particular target variable. We agree completely with their idea of selecting tuning parameters using forecast loss criteria, when the objective to prediction of a particular target variable.

© 2016 American Statistical Association
Journal of Business & Economic Statistics
July 2016, Vol. 34, No. 03
DOI: 10.1080/07350015.2016.1186554

In a third part of their article, the authors address an important characteristic of economic datasets, namely, that of instability. In a series of well-formulated arguments that are corroborated with a direct empirical evidence, they show that the dimension reduction methods that they examine broadly deliver robust predictions, and hence are useful for guarding against instability. In particular, they carry out the fluctuation test due to Giacomini and Rossi (2010) and the robust rationality test due to Rossi and Sekhposyan (2014). Using these tests, they establish that the superior predictive performance of their "big data" methods relative to that of an autoregressive strawman model remains constant, and that associated predictions are rational. Taken together, these findings constitute evidence of the robustness to instability of the data reduction methods examined in this article.

3. QUESTIONS AND DIRECTIONS

3.1 Alternative Methods for Uncovering Latent Factors

In CR, principal component analysis (PCA) is implemented to estimate latent factors. (The PCA estimator used in CR is based on solving the least-square estimation problem discussed in Stock and Watson (2002). In the sequel, we focus our discussion on this estimator. For a discussion of one important alternative estimator based on a dynamic factor model, see Forni et al. (2000), where principal components are extracted from the frequency domain.) PCA yields uncorrelated latent principal components (i.e., diffusion indexes) via the use of data projection in the direction of the maximum variance; and principal components (PCs) are naturally ordered in terms of their variance contribution. The first PC defines the direction that captures the maximum variance possible, the second PC defines the direction of maximum variance in the remaining orthogonal subspace, and so forth. As discussed above, one of the key contributions of CR is their discussion of methods for targeting PCs for use in forecasting based not on total variance explanatory power, but based on the usefulness in predicting a particular variable. In the next subsection, we shall discuss this issue in more detail.

Perhaps because the derivation of PCs in PCA is easily done via use of singular value decompositions, it is one of the most frequently used methods for constructing diffusion indexes (see, e.g., Bai and Ng 2002, 2006; Stock and Watson 2002 for details). In Kim and Swanson (in press), two additional methods for estimating latent factors are examined in the context of forecasting, including independent components analysis (ICA) and sparse principal components analysis (SPCA). ICA has previously been used in economics for macroeconomic forecasting by Moneta et al. (2013), Tan and Zhang (2012), and Yau (2004). However, SPCA appears to have been largely ignored, to date. ICA (see, e.g., Comon 1994; Lee 1998) uses differential entropy or "negentropy" as a measure of non-Gaussianity to construct statistically independent factors with non-Gaussian distributions. This method is closely related to projection pursuit, because it is assumed that the most interesting projections of multivariate datasets are those that show the least Gaussianity. SPCA is designed to uncover *uncorrelated* components and ultimately factors, just like PCA. However, the method also searches for components whose factor loading coefficient matrices are "sparse" (i.e., factor loading can be identically zero). This is different from PCA, which imposes nonzero loadings for the entire set of variables. In this sense, SPCA can deliver more parsimonious latent factors than PCA or ICA (for further discussion, see Vines 2000; Jolliffe, Trendafilov, and Uddin 2003; Zou, Hastie, and Tibshirani 2006).(SPCA is widely used in other disciplines, such as in gene expression genomics, see, e.g., Carvahlo et al. 2008.)

An interesting question to ask in empirical contexts is whether ICA and SPCA are useful alternatives to PCA, when carrying out forecasting experiments. ICA, for example, has been found to be useful in acoustic and image signal extraction problems (see, e.g., Hyvärinen and Oja 2000). What about economic applications such as signal extraction in high frequency finance applications, or economic forecasting?

One potential advantage of SPCA is the sparseness feature. It is well accepted that parsimonious models perform well for forecasting. Does this stylized fact apply also to the construction of diffusion indexes? Also, does the imposition of sparseness make the economic interpretation of diffusion indexes easier by opening up the "black box"? (See Armah and Swanson (2010a, 2010b) for further discussion of observable diffusion index proxies and the benefits of parsimonious factor augmented forecasting models.)

Note that ridge regression, as discussed in CR, is a penalized regression method in which the estimator, say $\widehat{\theta}$, is obtained via solving the following problem:

$$\widehat{\theta}_{\text{ridge}} = \arg\min_{\theta} \left\| y - \Sigma_{i=1}^{N} X_i \theta_i \right\|^2 + \lambda \Sigma_{i=1}^{N} \theta_i^2,$$

where y is the $Tx1$ target variable, $X = [X_1, \ldots, X_N]$, $i = 1, \ldots, N$ is the TxN predictor matrix, with $X_i = (X_{1,i}, \ldots, X_{T,i})'$, and $\lambda > 0$ is the tuning parameter. Alternative estimators that may be worth considering in the context of CR's analysis include the lasso and the elastic net, formulated as

$$\widehat{\theta}_{\text{lasso}} = \arg\min_{\theta} \left\| y - \Sigma_{i=1}^{N} X_i \theta_i \right\|^2 + \lambda \Sigma_{i=1}^{N} \left| \theta_i \right|,$$

and

$$\widehat{\theta}_{\text{elastic net}} = (1 + \lambda_2) \left\{ \arg\min_{\theta} \left\| y - \Sigma_{i=1}^{N} X_i \theta_i \right\|^2 + \lambda_1 \Sigma_{j=1}^{N} \left| \theta_j \right| + \lambda_2 \Sigma_{j=1}^{N} \theta_j^2 \right\}.$$

Interestingly, SPCA is introduced in Zou, Hastie, and Tibshirani (2006) by first formulating PCA as a regression-type optimization problem, and then by subsequently imposing lasso (elastic net) constraints on the regression coefficients in the optimization problem. In this sense, SPCA is a natural data reduction method to examine in contexts where penalized regression is being used, and is a natural alternative to the PCA method used by CR.

As an illustration of the forecasting trade-offs associated with using PCA, ICA, and SPCA, please refer to Table 1. The results in this table are based on the set of experiments reported in Kim and Swanson (KS: in press), although they are not presented in that article. In KS, $h = 1$, 3, and 12 month ahead predictions of 11 macroeconomic variables from the Stock and Watson (2002, 2012) dataset are constructed, using a predictor matrix including all 144 variables in that dataset. The 11 target variables to

Table 1. Summary of MSFE "best" factor estimation methods

Panel A: Recursive window estimation

Specification	Horizon	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
SP1	$h = 1$	PCA	SPCA	SPCA	ICA	ICA	SPCA	SPCA	SPCA	PCA	SPCA	ALL
	$h = 3$	PCA	PCA	ICA	SPCA	PCA	SPCA	PCA	SPCA	SPCA	PCA	SPCA
	$h = 12$	SPCA	SPCA	SPCA	PCA	PCA	SPCA	PCA	PCA	SPCA	SPCA	ICA
SP1L	$h = 1$	PCA	PCA	PCA	PCA	PCA	PCA	ALL	PCA	PCA	PCA	PCA
	$h = 3$	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
	$h = 12$	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	ICA
SP2	$h = 1$	PCA	PCA	PCA	PCA	ICA	PCA	ALL	PCA	SPCA	PCA	ICA
	$h = 3$	PCA	PCA	PCA	PCA	ALL	PCA	PCA	PCA	SPCA	SPCA	PCA
	$h = 12$	SPCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	SPCA	PCA	PCA
SP2L	$h = 1$	PCA	PCA	PCA	ICA	SPCA	PCA	ALL	PCA	PCA	PCA	ICA
	$h = 3$	PCA	PCA	PCA	PCA	ALL	ICA	PCA	SPCA	PCA	SPCA	PCA
	$h = 12$	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	SPCA	PCA

Panel B: Rolling window estimation

Specification	Horizon	UR	PI	TB10Y	CPI	PPI	NPE	HS	IPX	M2	SNP	GDP
SP1	$h = 1$	PCA	PCA	PCA	PCA	PCA	PCA	ALL	PCA	PCA	PCA	PCA
	$h = 3$	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
	$h = 12$	SPCA	PCA	PCA	SPCA	PCA	SPCA	SPCA	PCA	PCA	ICA	SPCA
SP1L	$h = 1$	PCA	PCA	PCA	PCA	PCA	PCA	ALL	PCA	PCA	PCA	ICA
	$h = 3$	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
	$h = 12$	PCA	PCA	SPCA	SPCA	PCA	ICA	SPCA	PCA	PCA	PCA	SPCA
SP2	$h = 1$	PCA	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	SPCA	PCA	PCA
	$h = 3$	PCA	PCA	ICA	SPCA	ICA	PCA	PCA	PCA	PCA	PCA	PCA
	$h = 12$	PCA	PCA	PCA	PCA	PCA	ICA	PCA	PCA	SPCA	SPCA	SPCA
SP2L	$h = 1$	PCA	PCA	PCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA
	$h = 3$	ICA	PCA	PCA	SPCA	SPCA	PCA	PCA	PCA	PCA	PCA	PCA
	$h = 12$	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	PCA	SPCA

Panel C: Summary of MSFE-best by PC method

	Recursive window estimation									Rolling window estimation								
	$h = 1$			$h = 3$			$h = 12$			$h = 1$			$h = 3$			$h = 12$		
	PCA	ICA	SPCA	PCA	ICA	SPCA	PCA	ICA	SPCA	PCA	ICA	SPCA	PCA	ICA	SPCA	PCA	ICA	SPCA
SP1	2	2	6	5	1	5	4	1	6	10	0	0	10	0	1	5	1	5
SP1L	10	0	0	10	0	1	10	1	0	9	1	0	11	0	0	6	1	4
SP2	7	2	1	8	0	2	8	0	3	9	0	2	8	2	1	7	1	3
SP2L	7	2	1	7	1	2	9	0	2	10	0	1	8	1	2	10	0	1

NOTES: See above discussion for an explanation of specification types used in this table, and well as a discussion of PCA, ICA, and SPCA. Results are based on comparing mean square forecast errors (MSFEs) across a number of benchmark linear models and data-reduction methods including PCA and RR, among others. The prediction period is 1974 through 2009 (see above and Kim and Swanson 2016 for complete details).

be forecasted are: the unemployment rate (UR), personal income (PI), the 10-year Treasury bond rate (TB), the CPI, the PPI, nonfarm payroll employment (NPE), housing starts (HS), IP, M2, the S&P500 (SNP), and GDP. Data are transformed to stationarity before prediction models are estimated, and models include a wide variety of benchmark linear models as well as penalized regression, machine learning, and shrinkage methods, including: bagging, boosting, ridge regression, least angle regression, elastic net, and the nonnegative garotte (see KS for further details). Various specification methods are considered using these methods, as discussed in KS, including:

Specification Type 1: Factors are constructed using PCA, ICA, and SPCA; and then prediction models are formed using the above methods to select functions of and weights for the factors to be used in prediction models of the variety given by

$$y_{t+h} = W_t \beta_W + F_t \beta_F + \varepsilon_{t+h}, \tag{1}$$

where h is the forecast horizon, W_t is a $1 \times s$ vector (possibly including lags of y), and F_t is a $1 \times r$ vector of factors. The parameters, β_W and β_F are defined conformably, and ε_{t+h} is a disturbance term. This specification type is estimated with and without lags of factors, denoted by SP1 and SP1L, respectively.

Specification Type 2: Factors are constructed using subsets of variables from the large-scale dataset and each of PCA, ICA, and SPCA. In particular, variables used in factor calculations are preselected via application of shrinkage methods. Thereafter, prediction models as above are estimated. This is different from the above approach of estimating factors using all of the variables. Note that forecasting models are again estimated with and without lags of factors, denoted by SP2 and SP2L, respectively.

The dataset used for the prediction experiment contains data from 1960 through 2009, and the prediction period runs from 1974:3 through 2009 (see KS for complete details). Inspection of the results in this table indicates that PCA often yields mean square forecast error (MSFE) "best" predictions, although this is not always the case. More specifically, SPCA (and to a lesser extent ICA) is "preferred" to PCA when the forecasting horizon is 1 month, while PCA dominates at all longer horizons. Is it possible that instability is driving these results?

3.2 Using Targeted Diffusion Indexes to Predict Select Variables

When targeting a selected variable for prediction, CR rightly stress that simply choosing the maximal eigenvalue diffusion indexes (i.e., those that explain the maximum variance possible, across the entire large-scale dataset) may not necessarily be an approach that should be expected to dominate use of other latent factors. Indeed, latent factors that are less important in explaining the variability across *all* variables in the dataset may actually be *very* important for predicting *particular* variables.

Mallow's C_L and the GCV criteria discussed above are analyzed by CR in the context of PCA (in which case the number of factors is chosen using either C_L or GCV). For other data reduction methods, including Ridge, LF, and PLS, these criteria are also implemented, although for PLS leave-one-out-cross-validation is instead used (see article for further details). A key difference between the prediction approach implemented in CR and the approach of Onatski (2015), for example, is that Onatski analyzes the asymptotic properties of the squared estimation error, $\left[\left(\widehat{\Lambda}\widehat{F}_t' - \Lambda F_t'\right)\left(\widehat{\Lambda}\widehat{F}_t' - \Lambda F_t'\right)'\right]/NT$ while CR analyze the mean square prediction error of the target variable, y (i.e., they analyze $\left\|X\widehat{\delta}^{\alpha} - X\delta\right\|^2$). This is an important difference, and underscores the fact that rankings associated with asymptotic approximations are dependent upon what the objective function is. When undertaking prediction experiments, CR rightly stress the importance of focusing on the target variable of interest, and do not rank methods based on the quality of the approximation of the diffusion index. In general, two obvious ways to specify diffusion indexes are as follows.

- First, one can simply implement standard PCA analysis, and use the k "highest variance contribution" diffusion indexes in subsequent forecasting. This approach has been used countless times in the literature, in cases where r is estimated using a variety of different statistics (see, e.g., Bai and Ng 2002; Onatski 2009, 2010; Ahn and Horenstein 2013; Li, Li, and Shi 2016, and the references cited therein).
- Second, after constructing the entire set of orthogonal diffusion indexes, one can *select* among them using the quality of out-of-sample predictions as a determining measure.

Of note is that when using the approach outlined in (ii), there is no reason to preselect the maximum number of diffusion indexes, say $r_{\max}$, via the use of one of the statistics from the articles cited in (i) above. Instead, one might consider even *low eigenvalue* diffusion indexes. In this context, CR show that the use of their criteria (including C_L and GCV) yields substantial prediction improvements when there are many diffusion indexes or when indexes specified in the standard way are not related to the target variable. (When implementing C_L and GCV, CR estimate the number of factors only once, in their first estimation window. This approach has clear computational advantages, although further research into more flexible estimation schemes warrants additional research.) Another approach in this context involves examining a *large* set of diffusion indexes via use of penalized regression methods. Namely, one can use penalized regression, coupled with PCA, to select a set of k orthogonal diffusion indexes that are targeted to a particular variable. Of course, this does not address the always present issue of rotation invariance. However, it is an obvious way to link minimization of prediction loss to diffusion index selection. In Kim and Swanson (2014), this idea is partially implemented, as they select a subset of k^* indexes from among the initial set of k indexes using various machine learning, variable reduction, and shrinkage methods, where the initial set of k indexes is preselected using the standard statistics in (i). Given that SPCA involves using the lasso (elastic net) to modify diffusion indexes constructed using PCA, it is clearly a method worth considering in this context. In summary, there are many methods available for data dimension reduction that deserve further exploration in the context of economic forecasting.

Another approach to targeting diffusion indexes is to preselect a subset of the N variables using the approaches discussed by Bai and Ng (2008, 2009), and Kim and Swanson (2014, in press). There are at least two reasons for considering subsets of variables when constructing diffusion indexes. First, a natural way to drill down to a single target variable to be predicted via determination of which subset(s) of variable(s) are most useful for prediction of said variable. Second, consider the problem of constructing diffusion indexes to explain variables in a financial dataset containing returns on the firms in the S&P500. It is not hard to imagine that idiosyncratic components common within industry groups are not common across all firms. Moreover, when estimating diffusion indexes based on the entire set of variables, this feature might act to "confound" estimates of factor loadings. For this reason, it may be of interest to attempt to estimate different types of diffusion indexes, including market specific, industry specific, and firm specific, say. In this context, variable subset analysis may prove crucial. For a related discussion of this topic, in the context of approximate factor models based on mixed frequency data, see Andreou et al. (2016).

3.3 Instability

As discussed above, SPCA seems to perform well at a one-step ahead horizon, while PCA performs well at longer horizons. Why is this? Is it possible that a subset of the loadings that are shrunk to zero using SPCA are those that are "best" at signaling structural change? This may certainly be true, even if the additional variance explained by inclusion of the variables associated with these loading is very small. Put differently, while parsimony might be useful when minimizing forecast loss (e.g., MSFE), and in diffusion index interpretation, it may also be the case that parsimony reduces the ability of a diffusion index to "adapt" to structural change. Along these lines, and as noted in Stock and Watson (2009), PCA-type diffusion indexes (in which no loadings are zero) may in some cases play an averaging or pooling role (similar to forecast combination) that leads to robustness against mild instability. There are clearly many pros and cons to the use of SPCA, which remain to be fully investigated. One clear take-away from the above discussion is that squared error loss (or variance reduction) may not be a natural measure for diffusion index selection (or loading estimation) in prediction contexts—say if interest focuses on turning point predictability. Indeed, hybrid loss functions that include a square error loss component as well as a turning point component may be more useful when constructing and selecting diffusion indexes in forecasting contexts in which predictions during transitions between recession and expansion are particularly important.

CR examine the robustness to instability of their dimension reduction methods, relative to the stability of a strawman autoregressive benchmark prediction model. They do this via the use of the Giacomini and Rossi (2010) fluctuation test (i.e., by testing whether predictions based on data reduction methods perform "better" over time than analogous forecasts constructed using their strawman model). They also check for robustness using the related rationality test of Rossi and Sekhposyan (2015) that is robust to instabilities. As discussed above, they find that (i) various of their proposed data reduction methods that are used to construct forecasts based on GCV are "better" than those based on traditional methods, and (ii) rationality across time is often achieved. Both of these results suggest that their methods are useful for "guarding" against instability. This remains an important area for research. Advances in this area other than the two articles just cited include those made by Breitung and Eickmeier (2011) and Corradi and Swanson (2014), who develop tests for the constancy of factor loadings, and the joint constancy of factor loadings and forecast regression coefficients, respectively. Early research into methods for tracking or monitoring structural change is discussed in Chu, Stinchcombe, and White (1996). The development of tests and methods that "signal" when the magnitude of structural change exhibited in a model generates instability that is beyond the ability of standard robust methods such as those discussed in CR to handle remains an interesting topic for future research, however.

ACKNOWLEDGMENTS

This comment derives from the discussion of the above article given at the 2016 Annual Allied Social Sciences Association meeting, special Journal of Business and Economics session sponsored by the American Statistical Association.

References

Ahn, S., and Horenstein, A. (2013), "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81, 1203–1227. [351]

Andreou, E., Gagliardini, P., Ghysels, E., and Rubin, M. (2016), "Is Industrial Production Still the Dominant Factor for the US Economy?" Working Paper, University of North Carolina, Chapel Hill. [351] Q13

Armah, N. A., and Swanson, N. R. (2010a), "Seeing Inside the Black Box: Using Diffusion Index Methodology to Construct Factor Proxies in Largescale Macroeconomic Time Series Environments," *Econometric Reviews*, 29, 476–510. [349]

——— (2010b), "Diffusion Index Models and Index Proxies: Recent Results and New Directions," *European Journal of Pure and Applied Mathematics*, 3, 478–501. [349]

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [348,349,351]

——— (2006), "Evaluating Latent and Observed Factors in Macroeconomics and Finance," *Journal of Econometrics*, 131, 507–537. [349]

——— (2008), "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, 146, 304–317. [351]

——— (2009), "Boosting Diffusion Indices," *Journal of Applied Econometrics*, 24, 607–629. [351]

Breitung, J., and Eickmeier, S. (2011), "Testing for Structural Breaks in Dynamic Factor Models," *Journal of Econometrics*, 163, 71–84. [352]

Carrasco, M., and Rossi, B. (2016), "In-Sample Inference and Forecasting in Misspecified Factor Models," *Journal of Business and Economic Statistics*. [xxxx] Q14

Carvahlo, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008), "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics," *Journal of the American Statistical Association*, 103, 1438–1456. [349]

Chu, J. C.-S., Stinchcombe, M., and White, H. (1996), "Monitoring Structural Change," *Econometrica*, 64, 1045–1065. [352]

Comon, P. (1994), "Independent Component Analysis—A New Concept?" *Signal Processing*, 36, 287–314. [349]

Corradi, V., and Swanson, N. R. (2014), "Testing for Structural Stability of Factor Augmented Forecasting Models," *Journal of Econometrics*, 182, 100–118. [352]

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000), "The Generalized Dynamic-Factor Model: Identification and Estimation," *Review of Economics and Statistics*, 82, 540–554. [349]

Giacomini, R., and Rossi, B. (2010), "Forecast Comparison in Unstable Environments," *Journal of Applied Econometrics*, 25, 595–620. [349,352]

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. (1987), *Discovering Causal Structure*, San Diego, CA: Academic Press. [348]

Hyvärinen, A., and Oja, E. (2000), "Independent Component Analysis: Algorithms and Applications," *Neural Networks*, 13, 411–430. [349]

Jolliffe, I., Trendafilov, N., and Uddin, M. (2003), "A Modified Principal Component Technique Based on the Lasso," *Journal of Computational and Graphical Statistics*, 12, 531–547. [349]

Kim, H.-H., and Swanson, N. R. (2014), "Forecasting Financial and Macroeconomic Variables Using Data Reduction Methods: New Empirical Evidence," *Journal of Econometrics*, 178, 352–367. [351]

——— (in press), "Mining Big Data Using Parsimonious Factor, Machine Learning, Variable Selection, and Shrinkage Methods," *International Journal of Forecasting*. [349] Q15

Lee, T.-W. (1998), *Independent Component Analysis—Theory and Applications*, Boston MA, : Springer. [349]

Li, H., Li, Q., and Shi, Y. (2016), "Determining the Number of Factors when the Number of Factors Can Increase With Sample Size," Working Paper, Texas A&M University . [351] Q16

Moneta, A., Entner, D., Hoyer, P., and Coad, A. (2013), "Causal Inference by Independent Component Analysis with Applications to Micro- and Macroeconomic Data," *Oxford Bulletin of Economics and Statistics*, 75, 705–730. [349]

Onatski, A. (2009), "Testing Hypotheses About the Number of Factors in Large Factor Models," *Econometrica*, 77, 1447–1479. [351]

——— (2010), "Determining the Number of Factors From Empirical Distribution of Eigenvalues," *Review of Economics and Statistics*, 92, 1004–1016. [xxxx]

——— (2015), "Asymptotic Analysis of the Squared Estimation Error in Misspecified Factor Models," *Journal of Econometrics*, 186, 388–406. [351]

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press. [348]

Pearl, J., and Verma, T. S. (1994), "A Theory of Inferred Causation," in *Logic, Methodology and Philosophy of Science IX*, eds. D. Prawitz and D. Westerstahl, Holland: Elsevier, pp. 789–811. [348]

Rossi, B., and Sekhposyan, T. (2015), "Forecast Rationality Tests in the Presence of Instabilities," *Journal of Applied Econometrics*. [352] Q17

Simon, H. (1955), "Causality and Econometrics: A Comment," *Econometrica*, 23, 193–195. [348]

Spearman, C. (1904), "General Intelligence Objectively Determined and Measured," *American Journal of Psychology*, 15, 201–293. [348]

Stock, J. H., and Watson, M. W. (2002), "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179. [349]

——— (2009), "Forecasting in Dynamic Factor Models Subject to Structural Instability," in *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*, eds. J. Castle and N. Shephard, Oxford: Oxford University Press, pp. 1–57. [352]

——— (2012), "Generalized Shrinkage Methods for Forecasting Using Many Predictors," *Journal of Business and Economic Statistics*, 30, 481–493. [349]

Swanson, N. R., and Granger, C. W. J. (1997), "Impulse Response Functions Based on a Causal Approach to Residual Orthogonalization in Vector Autoregressions," *Journal of the American Statistical Association*, 92, 357–367. [348]

Tan, L., and Zhang, H. (2012), "Forecast of Employment Based on Independent Component Analysis," in *Information Computing and Applications - Proceedings, Third International Conference*, eds. C. Liu, L. Wang, and A. Yang, Berlin: Springer-Verlag, pp. 373–381. [349]

Vines, S. K. (2000), "Simple Principal Components," *Applied Statistics*, 49, 441–451. [349]

Wold, H. (1954), "Causality and Econometrics," *Econometrica*, 22, 162–177. [348]

Wright, S. (1921), "Correlation and Causation," *Journal of Agricultural Research*, 20, 557–585. [348]

Yau, R. (2004), "Macroeconomic Forecasting with Independent Component Analysis," Econometric Society 2004 Far Eastern Meetings Monograph no. 741, Econometric Society. [349]

Zou, H., Hastie, T., and Tibshirani, R. (2006), "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, 15, 262–286. [349]

Rejoinder: In-Sample Inference and Forecasting in Misspecified Factor Models

Marine Carrasco
Departement de Sciences Economiques, Université de, Montréal, CP 6128, succ Centre Ville, Montreal, CIREQ, CIRANO, QC H3C3J7, Canada (*marine.carrasco@umontreal.ca*)

Barbara Rossi
ICREA-Univ., Pompeu Fabra, Barcelona GSE and CREI, Universitat Pompeu Fabra, C/Ramon Trias Fargas 25-27, 08005, Barcelona, Spain (*barbara.rossi@upf.edu*)

1. INTRODUCTION

We thank all the discussants for their constructive comments and stimulating discussions of our article. Each of them raised interesting and insightful issues on various themes, which we discuss in the following sections.

2. ALTERNATIVE WAYS TO EXTRACT INFORMATION FROM A LARGE DATASET OF PREDICTORS

In particular, the careful Monte Carlo simulation analysis undertaken in Xu Cheng and Bruce Hansen's insightful comment complements our own results by providing a comparison with the method that they proposed. As pointed out in James Stock's comment, our approach might look similar to Xu Cheng and Bruce Hansen's approach, since both papers rely on cross-validation. However, there are some major differences. While we use generalized cross-validation (GCV) and Mallows criteria on raw data to select a regularization parameter, they apply Mallows and Leave-h-out cross-validation to select weights in a model combination, where each model is a factor-augmented regression. Our method does not assume a factor model.

We recognize that there are several other methods that we did not include in our analysis, such as those considered in the insightful comment by Norman Swanson. In our article, we have considered GCV and Mallows as possible criteria to choose tuning parameters, and it would be an interesting avenue of research to evaluate whether it could be advantageous to implement these criteria in other techniques such as those discussed in his comment.

3. THE IMPORTANCE OF THE CHOICE OF TUNING PARAMETERS

All of these techniques require a tuning parameter. We agree with James Stock that it is important to carefully justify the choice of the tuning parameters: this choice needs to be free from data-mining to evaluate actual out-of-sample predictive ability. The particular choice we made for the value of d in our implementation of Landweber–Fridman (LF) is guided by the value used in an unrelated Monte Carlo simulation in a previous data-reduction article (Carrasco and Noumon 2012). Note that, in our empirical application, the value of the other tuning parameter in Landweber–Fridman (α) has been chosen using the part of the sample that does not include the out-of-sample forecasting portion of the data, and hence it is free of data mining. Note that the parameter d could also be chosen via cross-validation over both d and α, although that would be computationally more

*M. Carrasco gratefully acknowledges partial financial support from SSHRC. B. Rossi gratefully acknowledges financial support from the ERC Grant #615608 and the Spanish Ministry of Economy and Competitiveness Grant ECO2012-33247.

© 2016 American Statistical Association
Journal of Business & Economic Statistics
July 2016, Vol. 34, No. 3
DOI: 10.1080/07350015.2016.1191500
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jbes*.