TI 2011-122/4
Tinbergen Institute Discussion Paper

Sparse and Robust Factor Modelling

Christophe Croux[1]
Peter Exterkate[2]

[1] *Faculty of Business and Economics, K.U. Leuven, Belgium;*
[2] *Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute.*

Tinbergen Institute is the graduate school and research institute in economics of Erasmus University Rotterdam, the University of Amsterdam and VU University Amsterdam.

More TI discussion papers can be downloaded at http://www.tinbergen.nl

Tinbergen Institute has two locations:

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Duisenberg school of finance is a collaboration of the Dutch financial sector and universities, with the ambition to support innovative research and offer top quality academic education in core areas of finance.

DSF research papers can be downloaded at: http://www.dsf.nl/

Duisenberg school of finance
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 8579

Sparse and Robust Factor Modelling

Christophe Croux*
Faculty of Business and Economics
K.U.Leuven

Peter Exterkate
Erasmus School of Economics
Erasmus University Rotterdam

July 25, 2011

Abstract

Factor construction methods are widely used to summarize a large panel of variables by means of a relatively small number of representative factors. We propose a novel factor construction procedure that enjoys the properties of robustness to outliers and of sparsity; that is, having relatively few nonzero factor loadings. Compared to more traditional factor construction methods, we find that this procedure leads to better interpretable factors and to a favorable forecasting performance, both in a Monte Carlo experiment and in two empirical applications to large data sets, one from macroeconomics and one from microeconomics.

Keywords: dimension reduction, forecasting, outliers, regularization.
JEL Classification: C38, C51, C53.

*Corresponding author. Address: Faculty of Business and Economics, K.U. Leuven, Naamsestraat 69, B-3000 Leuven, Belgium; email: christophe.croux@econ.kuleuven.be; phone: +32-16-326958; fax: +32-16-326732.

1 Introduction

Empirical researchers in a wide variety of fields face the problem of summarizing large data sets by a small number of representative factors, which can then be used for either descriptive or predictive purposes. In particular, the econometrics literature of the last decade contains successful applications of factor models to forecasting macroeconomic time series (Stock and Watson, 2002; Bai and Ng, 2008) and excess returns in stock and bond markets (Ludvigson and Ng, 2007, 2009).

Principal component analysis (PCA) is the classical tool for extracting such factors. In recent years, however, two major drawbacks of PCA have received attention. First, PCA lacks robustness to outliers. Even a very small proportion of data contamination results in inaccurate factors. This problem has been alleviated by explicitly downweighting such observations (Pison et al., 2003), by employing more robust loss functions than the usual sum of squares (De la Torre and Black, 2001), or by a combination of both approaches (Croux et al., 2003; Maronna and Yohai, 2008).

Second, in standard PCA all variables generally load on all extracted factors; that is, every original variable is represented as a linear combination of all factors. This feature leads to difficulties in giving an interpretation to the factors, as well as to a loss of degrees of freedom and thus to unnecessarily large estimation uncertainties. Penalized variants of standard PCA to overcome this problem have recently been developed by Jolliffe et al. (2003) and Witten et al. (2009), among others.

In this paper, we propose a factor construction method that unifies both approaches, yielding robust factors with sparse loadings. Our procedure is a combination of the robust estimation methods from Maronna and Yohai (2008) and the penalization technique introduced by Witten et al. (2009). We provide a relatively simple alternating algorithm to solve the resulting optimization problem, and we document the good interpretability and forecasting properties of our method in a Monte Carlo study and in two empirical applications. Our first application concerns forecasting key U.S. macroeconomic variables, as in Stock and Watson (2002). The other application is microeconomic: we analyze the Boston housing data set from Harrison and Rubinfeld (1978). The results show that ignoring the presence of outlying observations, which are often overlooked in empirical econometric studies, has important consequences for forecast accuracy.

To the best of our knowledge, our proposed method is the first to combine robustness and sparsity in the

context of factor modelling. Moreover, while factors models are common in the macroeconomic forecasting literature, robustness issues are typically only considered in small sets of predictors (Fagiolo et al., 2008; Bańbura et al., 2010). Sparsity is not commonly studied either, although a related approach using reduced-rank vector autoregressions was recently found to improve macroeconomic forecasts by Carriero et al. (2011).

The remainder of this article is structured as follows. We describe the methodology in Section 2 and test it in a simulation study in Section 3. Empirical applications to macroeconomic forecasting and to the Boston housing data set follow in Sections 4 and 5, respectively, and Section 6 concludes.

2 Methodology

2.1 Robust Matrix Approximation

We consider the problem of approximating an $n \times p$ matrix X by a rank-q matrix $\hat{X} = FA'$, where F has dimensions $n \times q$ and A is $p \times q$. The standard way to proceed is to apply principal component analysis (PCA), in which F and A are estimated by minimizing

$$Q_{L_2}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \sum_{i=1}^{n} \left(x_{ij} - f_i' a_j\right)^2, \tag{1}$$

where f_i and a_j denote rows of F and A, respectively. Although it is well-known that Q_{L_2} can be minimized using the singular value decomposition of X, we note that an alternating approach (due to Wold, 1966) is also possible. Given initial estimated of F and A, we iterate until convergence:

- Solve (1) for A by solving p ordinary least-squares (OLS) problems: the jth row is $a_j = (F'F)^{-1} F'x_j$, where x_j denotes the jth column of X.
- Solve (1) for F by solving n OLS problems: the ith row is $f_i = (A'A)^{-1} A'x_i$, where x_i denotes the ith row of X.

As all least-squares procedures, PCA is very sensitive to outlying observations (Maronna et al., 2006). A more robust alternative to (1) is to replace the sums of squared deviations by sums of absolute deviations; that is, to minimize

$$Q_{L_1}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \sum_{i=1}^{n} \left| x_{ij} - f_i' a_j \right|. \tag{2}$$

This L_1 minimization problem can be solved using a similar alternating algorithm as in the L_2 case, replacing OLS regressions by least absolute deviations (LAD) regressions. This procedure was advocated by Croux et al. (2003), who labelled it Robust Alternating L_1 Regressions (RAR).

Maronna and Yohai (2008) propose to replace the squared or absolute deviations by an even more robust error measure, using the Tukey biweight loss function $\rho(r) = \min\left\{1, \left(1 - (r/c)^2\right)^3\right\}$. This loss function is bounded, which makes it very robust to large outliers. The constant c is fixed at 3.4437, so that 85% efficiency at the normal distribution is attained. Because the Tukey loss function downweights large residuals, it is essential that the columns are appropriately scaled to decide what "large" means. Thus, for every variable j, let $\hat{\sigma}_j$ denote an estimate of the scale of the residuals $x_{ij} - f_i' a_j$, for $i = 1, 2, \ldots, n$. Then, Maronna and Yohai (2008) propose to minimize

$$Q_{\text{Tukey}}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \hat{\sigma}_j^2 \sum_{i=1}^{n} \rho\left(\frac{x_{ij} - f_i' a_j}{\hat{\sigma}_j}\right). \tag{3}$$

As a robust scale estimate, they consider the median absolute deviation

$$\hat{\sigma}_j = 1.4826 \underset{i}{\text{median}} \left\{ \left| x_{ij} - f_i' a_j \right| \right\}. \tag{4}$$

If we would set $\rho(r) = r^2$, Criterion (3) would reduce to the PCA criterion (1). In order to be able to apply the alternating algorithm to minimize (3) for the Tukey loss function as well, we rewrite it as a weighted least squares (WLS) problem. Defining weights

$$w_{ij} = \left(\frac{x_{ij} - f_i' a_j}{\hat{\sigma}_j}\right)^{-2} \rho\left(\frac{x_{ij} - f_i' a_j}{\hat{\sigma}_j}\right), \tag{5}$$

the objective in equation (3) can be rewritten as

$$Q_{\text{Tukey}}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \sum_{i=1}^{n} w_{ij} \left(x_{ij} - f_i' a_j\right)^2. \tag{6}$$

This means that, given initial estimates of F, A, and the residual scales $\hat{\sigma}_j$, we can solve (3) by iterating the following scheme until convergence:

- Solve (6) for A by solving p WLS problems: the jth row is $a_j = (F'D_jF)^{-1} F'D_jx_j$, where D_j is a diagonal matrix containing $w_{1j}, w_{2j}, \ldots, w_{nj}$.
- Update $\hat{\sigma}_j$ for $j = 1, 2 \ldots, p$ using (4) and, hence, all weights w_{ij} using (5).
- Solve (6) for F by solving n WLS problems: the ith row is $f_i = (A'D_iA)^{-1} A'D_ix_i$, where D_i is a diagonal matrix containing $w_{i1}, w_{i2}, \ldots, w_{ip}$.
- Update the scale estimates $\hat{\sigma}_j$ and the weights w_{ij} again.

We shall consider all three different criteria introduced above. All columns of X are standardized before the estimation procedure. For the L_2 criterion (1) we standardize all columns to mean zero and variance one; for the L_1 criterion (2), to median zero and mean absolute deviation one; and for the Tukey criterion (3), to median zero and median absolute deviation one. Initial estimates for F and A are obtained as described by Maronna and Yohai (2008).

2.2 A Sparsity Condition

In factor-model terminology, the columns of F represent factors and A is the loading matrix. In order to improve the interpretability of the estimated factors, it may be desirable to impose a sparsity condition on the loading matrix; that is, to limit the number of nonzero factor loadings. In addition to improving interpretability, another interesting effect of such a condition is reducing the estimation uncertainty, which is an important consideration for forecasting. In the spirit of Witten et al. (2009), we implement this sparsity condition by adding an L_1 penalty to (1), (2), or (3): for some positive scalar λ, we aim to minimize

$$Q(F, A; X) + \lambda \sum_{j=1}^{p} \sum_{k=1}^{q} |a_{jk}|, \tag{7}$$

where Q denotes either Q_{L_2}, Q_{L_1}, or Q_{Tukey}. As it stands, objective (7) does not attain a minimum value. Although the linear subspace spanned by the columns of F is identified, we observe that for any candidate

minimum point $\left(\hat{F}, \hat{A}\right)$, the equivalent factorization $\left(c\hat{F}, \frac{1}{c}\hat{A}\right)$ leads to a smaller objective value for any $c > 1$. To remove this unwanted feature, we restrict the magnitude of F by adding another penalty term to (7). As our purpose is not to impose sparsity on F, this additional term will be an L_2 penalty: we minimize

$$Q\left(F, A; X\right) + \lambda \sum_{j=1}^{p} \sum_{k=1}^{q} |a_{jk}| + \nu \sum_{i=1}^{n} \sum_{k=1}^{q} f_{ik}^2. \tag{8}$$

Finally, we note that Problem (8) is overparameterized: if the factorization $\left(\hat{F}, \hat{A}\right)$ solves (8) for the penalty parameters (λ^*, ν^*), then the equivalent factorization $\left(c\hat{F}, \frac{1}{c}\hat{A}\right)$ is a solution for $\left(c\lambda^*, \frac{\nu^*}{c^2}\right)$ for any $c > 0$. Therefore, we lose no generality in fixing either λ or ν at a specific positive value. We set $\nu = 1/\left(2n\right)$, so that only λ measures the degree of sparsity.

The alternating procedures in Section 2.1 can be adapted for problem (8). First, given F and (in the Tukey case) the weights w_{ij}, finding the jth row of A amounts to minimizing

$$Q\left(F, A; X\right) + \lambda \sum_{k=1}^{q} |a_{jk}| . \tag{9}$$

For the L_2 and Tukey criterion functions, we recognize (9) as a Lasso problem (Tibshirani, 1996), with regressand $\left(\sqrt{w_{ij}}\right) x_{ij}$ and regressors $\left(\sqrt{w_{ij}}\right) f_i$. Efficient algorithms to solve this problem are known; see Friedman et al. (2010). For the L_1 criterion, minimizing (9) is a LAD-Lasso problem (Wang et al., 2007).

Second, given A (and the weights), finding the ith row of F is equivalent to minimizing

$$Q\left(F, A; X\right) + \frac{1}{2n} \sum_{k=1}^{q} f_{ik}^2. \tag{10}$$

The ridge regression problem (10) can be solved analytically for the L_2 and Tukey criteria, resulting in

$$f_i = \left(A' D_i A + I\right)^{-1} A' D_i x_i, \tag{11}$$

where we set $D_i = I$ in the L_2 case. For the L_1 criterion, we use a standard numerical minimization routine.

2.3 Tuning Parameters

The sparse and robust factor extraction procedure that we developed in Sections 2.1 and 2.2 is characterized by two tuning parameters; the number of factors (q) and the penalty parameter (λ). To specify values for q and λ, we minimize the Bayesian Information Criterion

$$BIC_{q,\lambda} = 2\sum_{j=1}^{p} \log \hat{\sigma}_{j;q,\lambda} + \mathrm{df}_{q,\lambda} \cdot \frac{\log n}{n}. \tag{12}$$

As argued by Zou et al. (2007), the "degrees of freedom" $\mathrm{df}_{q,\lambda}$ can be approximated by the number of nonzero entries in the estimated A. Further, we approximate the determinant of the residual covariance matrix by the product of scale estimates $\hat{\sigma}_j^2$, which are median absolute deviations (4) when using the Q_{Tukey} criterion, mean absolute deviations when using Q_{L_1}, and standard deviations when using Q_{L_2}. This amounts to discarding all covariances between columns of the residual matrix. We feel that this is a reasonable choice, as most of the correlation structure in X should be captured by the factors. Moreover, this procedure circumvents the nontrivial task of robustly estimating covariances.

3 Monte Carlo Simulation

To evaluate the potential of the sparse robust factor extraction procedure described in Section 2, we assess its performance through a Monte Carlo study. As $n \approx p$ is typical for situations to which factor modelling is applied, we simulate data sets with $n = p = 100$. The number of latent factors will be $q = 2$.

We generate data from a factor model $X = FA' + E$. Here, the matrix A contains the factor loadings, and we impose that its true structure is sparse. The loading matrix has 100 rows and two columns:

$$A = \begin{pmatrix} 10 \text{ rows} & (+1, & +1) \\ 10 \text{ rows} & (+1, & -1) \\ 10 \text{ rows} & (-1, & +1) \\ 10 \text{ rows} & (-1, & -1) \\ 60 \text{ rows} & (\ 0, & \ 0) \end{pmatrix}. \tag{13}$$

For the 100×2 matrix of latent factors F and the 100×100 matrix of noise E, we consider the following four data-generating processes:

- *Normal:* the entries of F and E are independent draws from the $N(0,1)$ distribution.
- *Heavy tails:* the entries of F are drawn from the $N(0,1)$ distribution, those of E from Student's t distribution with two degrees of freedom.
- *Vertical outliers:* like the "Normal" DGP, but a random selection of 10% of the entries of E are replaced by the value 20.
- *Bad leverage rows:* like the "Normal" DGP, but a random selection of 10% of the rows of F are replaced by $(+20, +40)$, and the corresponding rows of E are replaced by $(-20, -40)\, A'$.

Note the difference between the final two DGPs. If an observation is a vertical outlier, the latent factors behave normally but the observed variable is contaminated. On the other hand, in a bad leverage row the factors behave abnormally but the observed variables are not informative about this fact.

In Tables 1 and 2 we report average results over 1000 simulation runs for each of these DGPs. We consider the L_2, L_1, and Tukey loss functions. For each of these, we report results using both the unpenalized criteria (1)-(3) and the penalized criterion (8). In the latter case, the penalty parameter λ is selected by minimizing the BIC (12) over the grid $\{0.0001, 0.001, 0.01, 0.1, 1\}$. We treat the true number of factors ($q = 2$) as known.

Table 1 reports on the structure of the estimated loading matrix A. Specifically, it shows how many of the 60 zero rows and 40 nonzero rows of the true A were correctly identified as zero or nonzero. From these results, it is clear that unpenalized estimation methods cannot succeed in exactly estimating zero loadings. The results for all penalized methods, on the other hand, are quite good: the penalized L_1 criterion correctly estimates more than half of the zero rows. Moreover, except for the penalized L_2 criterion, there are no false zero rows in the estimated loading matrix; thus, all variables that load on the factors are correctly identified.

An important application of factor models is forecasting a variable y, which is assumed to be driven by (a subset of) the same factors that drive X; say, $y = F\beta + \eta$, where η is noise. After $\hat{F}$ is obtained as above, we would estimate β using a form of regression (either ordinary least squares or a more robust variant) on the observations for which y_i is known, and then construct a forecast $\hat{y}_i = \hat{f}_i'\hat{\beta}$ for the remaining observations.

Table 1: Estimated structure of the loading matrix in the Monte Carlo simulations.

		Number of rows				Number of rows	
		correct	correct			correct	correct
DGP	Criterion	zero	nonzero	DGP	Criterion	zero	nonzero
Normal	L_2, $\lambda = 0$	0	40	Vertical outliers	L_2, $\lambda = 0$	0	40
	L_2, $\lambda > 0$	8.781	40		L_2, $\lambda > 0$	11.957	34.872
	L_1, $\lambda = 0$	0	40		L_1, $\lambda = 0$	0	40
	L_1, $\lambda > 0$	27.326	40		L_1, $\lambda > 0$	37.977	40
	Tukey, $\lambda = 0$	0	40		Tukey, $\lambda = 0$	0	40
	Tukey, $\lambda > 0$	6.377	40		Tukey, $\lambda > 0$	6.995	40
Heavy tails	L_2, $\lambda = 0$	0	40	Bad leverage rows	L_2, $\lambda = 0$	0	40
	L_2, $\lambda > 0$	11.314	39.860		L_2, $\lambda > 0$	5.266	40
	L_1, $\lambda = 0$	0	40		L_1, $\lambda = 0$	0	40
	L_1, $\lambda > 0$	29.902	40		L_1, $\lambda > 0$	30.791	40
	Tukey, $\lambda = 0$	0	40		Tukey, $\lambda = 0$	0	40
	Tukey, $\lambda > 0$	5.710	40		Tukey, $\lambda > 0$	14.603	40

Notes: This table reports average results over 1000 replications of each of the data-generating processes described in the text. The numbers indicate how many of the rows of the loading matrix A were correctly estimated to be zero/nonzero; the true loading matrix contains 60 zero and 40 nonzero rows.

Instead of forecasting a specific linear combination of the factors, we consider the problem of forecasting *any* linear combination of the factors. The quality of such forecasts is assessed by computing the angle between the two-dimensional linear subspaces of $\mathbb{R}^{100}$ spanned by F and $\hat{F}$, respectively: the smaller this angle is, the more suitable $\hat{F}$ is for forecasting variables of the form $F\beta$.

The average values of this angle, again over 1000 simulation runs, are reported in the rightmost column of Table 2. Here, the value of using a penalized criterion function becomes apparent: in almost all cases, the angle between the true and estimated factors is smaller if a nonzero penalty is present. For the normal DGP, the different criterion functions yield similar results. For the other three DGPs, in which outliers are present, the L_2 factor estimates are much less accurate than the estimates obtained using more robust criterion functions. An extreme example is the "bad leverage rows" DGP, for which angles between 1.2 and 1.3 radians are observed. As a right angle measures $\pi/2 \approx 1.571$ radians, it is clear that the L_2 criterion is severely misguided by the bad leverage rows. The Tukey criterion performs remarkably well in this case.

We also report results for the approximation of the data matrix X in Table 2, expressed as the root mean squared error (RMSE), mean absolute error (MnAE), and median absolute error (MdAE). First, we notice

Table 2: Summary statistics for the Monte Carlo simulations.

		Approximation of X			Angle
DGP	Criterion	RMSE	MnAE	MdAE	$(F, \hat{F})$
Normal	$L_2, \lambda = 0$	**0.975**	0.778	0.658	0.225
	$L_2, \lambda > 0$	0.979	0.781	0.660	**0.219**
	$L_1, \lambda = 0$	0.991	**0.770**	**0.640**	0.259
	$L_1, \lambda > 0$	0.995	0.778	0.650	0.256
	Tukey, $\lambda = 0$	0.981	0.778	0.653	0.233
	Tukey, $\lambda > 0$	0.984	0.780	0.655	0.228
Heavy tails	$L_2, \lambda = 0$	**3.423**	1.478	0.915	0.435
	$L_2, \lambda > 0$	3.436	1.453	0.886	0.412
	$L_1, \lambda = 0$	3.487	**1.383**	**0.793**	0.295
	$L_1, \lambda > 0$	3.493	1.395	0.804	**0.291**
	Tukey, $\lambda = 0$	3.480	1.396	0.816	0.326
	Tukey, $\lambda > 0$	3.483	1.396	0.813	0.311
Vertical outliers	$L_2, \lambda = 0$	**5.873**	3.638	2.182	1.314
	$L_2, \lambda > 0$	5.898	3.599	2.147	1.332
	$L_1, \lambda = 0$	6.316	**2.681**	**0.747**	**0.286**
	$L_1, \lambda > 0$	6.325	2.697	0.763	0.288
	Tukey, $\lambda = 0$	6.312	2.692	0.762	0.300
	Tukey, $\lambda > 0$	6.315	2.694	0.764	0.291
Bad leverage rows	$L_2, \lambda = 0$	1.169	0.867	0.671	1.264
	$L_2, \lambda > 0$	1.185	0.880	0.697	1.289
	$L_1, \lambda = 0$	0.944	**0.701**	**0.562**	0.344
	$L_1, \lambda > 0$	0.948	0.706	0.569	0.338
	Tukey, $\lambda = 0$	**0.936**	0.708	0.575	0.325
	Tukey, $\lambda > 0$	0.938	0.713	0.577	**0.320**

Notes: This table reports average results over 1000 replications of each of the data-generating processes described in the text. In the group of columns headed "Approximation of X", X is compared to $\hat{X} = \hat{F}\hat{A}'$; the root mean squared error and the mean and median absolute error are reported. In the rightmost column, we report the angle between the linear subspaces spanned by the columns of F and $\hat{F}$, in radians; for the "Bad leverage rows" DGP, the bad leverage rows are removed for this computation. For each DGP, the smallest RMSE, MeanAE, MedianAE and angle are printed in boldface.

that the in-sample approximation of X is most accurate without a penalty term, and using the L_2 or L_1 loss function, depending on whether the approximation quality is measured in squared or absolute errors. This result was to be expected, as the corresponding objective minimizes this error. The only exception to this rule is the "bad leverage rows" DGP, where the L_2 algorithm apparently failed to converge. We also note that little accuracy is lost when a positive penalty term λ is applied, and that the differences between loss functions in RMSEs are minor. Measured in mean or median absolute errors, the differences between the results from using the Tukey or L_1 loss are still small, but L_2 performs markedly worse in all DGPs except the normal.

4 Application: Macroeconomic Forecasting

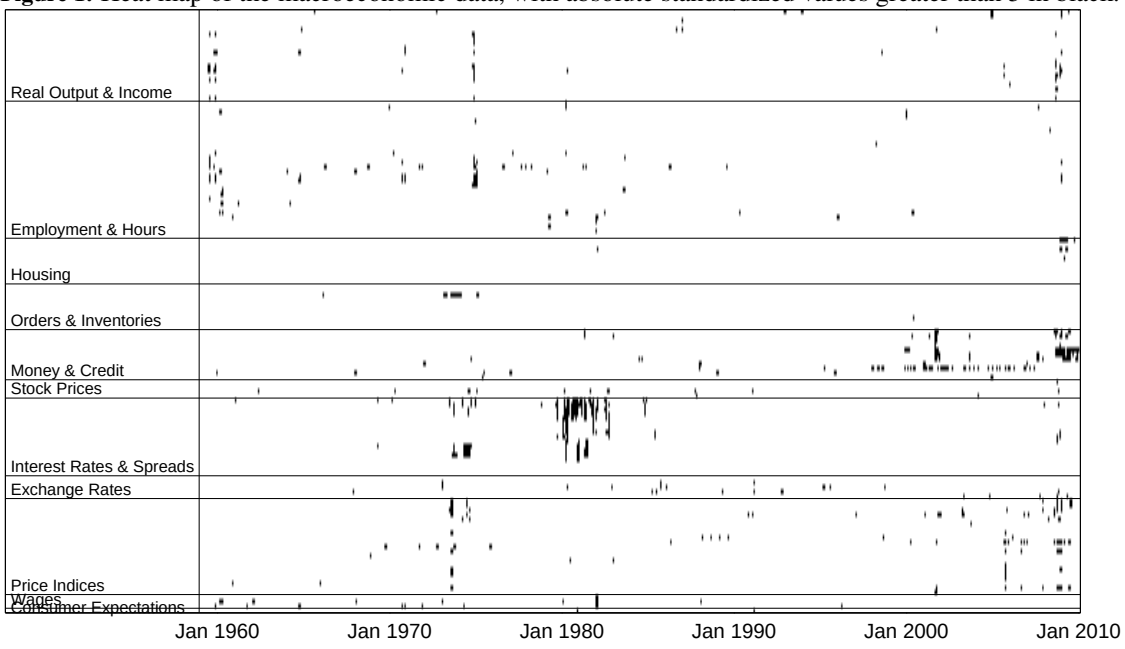
4.1 Data and Forecast Model

To evaluate the forecast performance of sparsely and robustly estimated factor models in an empirical application, we consider forecasting of four key macroeconomic variables. The data set consists of monthly observations on 132 U.S. macroeconomic variables, including various measures of production, consumption, income, sales, employment, monetary aggregates, prices, interest rates, and exchange rates. All series have been transformed to stationarity by taking logarithms and/or differences, as described in Stock and Watson (2002). They also define a partitioning of the data set into economically meaningful groups of related variables. We use an updated version of their data set, covering the period from January 1959 until (and including) January 2010, taken from Exterkate et al. (2011). Some of the 132 time series start later than January 1959, while a few other variables have been discontinued before the end of the sample period. For each month under consideration, observations on at most five variables are missing.

A heat map of this data set is shown in Figure 1. For this figure, all time series were standardized to have median zero and median absolute deviation one. Each entry of the resulting matrix is shown in either black or white, depending on whether the standardized value is greater or smaller than five in absolute value. Time runs along the horizontal axis, and the different time series are organized in groups of related variables shown along the vertical axis. Despite the efforts to transform the data to near normality, a relatively large number of outliers shows up in various time series, mainly in interest rates series during the monetarist experiment in 1979-82, and in money and credit series in the recessions of 2000-01 and (especially) 2008-09. For this reason, we analyze these data using the robust methods outlined in Section 2.

We focus on forecasting four key measures of real economic activity: Industrial Production, Personal Income, Manufacturing & Trade Sales, and Employment. (The acronyms by which Stock and Watson (2002) refer to these series are `ip`, `gmyxpq`, `msmtq`, and `lhnag`, respectively.) For each of these variables, we produce out-of-sample forecasts for the annualized h-month percentage growth rate, which are computed as $y_{t+h}^h = (1200/h) \ln(v_{t+h}/v_t)$, where v_t is the untransformed observation on the level of each variable in month t. We consider growth rate forecasts for $h = 1, 3, 6$ and 12 months.

Figure 1: Heat map of the macroeconomic data, with absolute standardized values greater than 5 in black.



The most widely used approach to forecasting in this setup is the diffusion index (DI) approach of Stock and Watson (2002), who document its good performance for forecasting these four macroeconomic variables. The DI methodology extends the standard principal component regression by including autoregressive lags as well as lags of the principal components in the forecast equation. Specifically, using ℓ_y autoregressive lags and ℓ_f lags of q factors, at time t, this "extended" principal-components method produces the forecast

$$\hat{y}^h_{t+h|t} = \hat{\alpha} + \sum_{s=0}^{\ell_y - 1} \hat{\beta}_s y^1_{t-s} + \sum_{s=0}^{\ell_f - 1} \sum_{k=1}^{q} \hat{\gamma}_{ks} \hat{f}_{k,t-s}. \tag{14}$$

The lags of the dependent variable in Equation (4.3) are one-month growth rates, irrespective of the forecast horizon h, because using h-month growth rates for $h > 1$ would lead to highly correlated regressors. In Stock and Watson (2002), the factors $\hat{f}_{kt}$ are standard principal components extracted from all 132 predictor variables, and $\hat{\alpha}$, $\hat{\beta}_s$ and $\hat{\gamma}_{ks}$ are OLS estimates.

In this study, we retain the forecast equation (4.3), but we change the estimation methods for the factors $\hat{f}_{kt}$ and the regression coefficients. In addition to standard principal components, which corresponds to the

L_2 criterion (1), we use the L_1 and Tukey variants of this criterion to estimate the factors. Moreover, we also estimate factors using the penalized criterion (8) for these three loss functions. After the $\hat{f}_{kt}$ have been obtained, we estimate the coefficient vector $\left(\alpha, \beta_0, \ldots, \beta_{\ell_y - 1}, \gamma_{10}, \ldots, \gamma_{q0}, \gamma_{11}, \ldots, \gamma_{q,\ell_f - 1}\right)'$ in (4.3) using either OLS, L_1 regression, or Tukey regression; the same loss function used to extract the factors. As the number of parameters is relatively small, we do not consider penalized regression estimation in this equation.

In each case, the lag lengths ℓ_y and ℓ_f, the number of factors q, and (if applicable) the penalty parameter λ are selected by minimizing the Bayesian Information Criterion (BIC). As our primary concern in this exercise is forecasting, we do not use expression (12) for the BIC, which measures how well the factors $\hat{F}$ fit X. Instead, we minimize

$$BIC_{\ell_y, \ell_f, q, \lambda} = 2 \log \hat{\sigma}_{\ell_y, \ell_f, q, \lambda} + (1 + \ell_y + \ell_f \cdot q) \frac{\log n}{n}, \tag{15}$$

where $(1 + \ell_y + \ell_f \cdot q)$ is the number of parameters in Equation (4.3), and where $\hat{\sigma}_{\ell_y, \ell_f, q, \lambda}$ is an estimate of the scale of the residuals $y^h_{t+h} - \hat{y}^h_{t+h|t}$. As in Section 2.1, this scale estimate is either the standard deviation, the mean absolute deviation, or the median absolute deviation, depending on which loss function is used.

As Stock and Watson (2002) find that allowing for multiple lags of the factors does not substantially improve the forecasting performance, we fix $\ell_f = 1$. For the other parameters, we allow $0 \leq \ell_y \leq 6$, $0 \leq q \leq 4$, and $\log_{10} \lambda \in \{-4, -3, -2, -1, 0\}$. Note that $\ell_y = 0$ and $q = 0$ correspond to using no autoregressive information and no information from factors, respectively.

All models are estimated on rolling windows with a fixed length of 120 months, such that the first forecast is produced for the growth rate during the first h months of 1970. For each window, the tuning parameter values are re-selected and the regression coefficients are re-estimated. That is, all of the tuning parameters (ℓ_y, q, λ) are allowed to differ over time and across methods.

4.2 In-Sample Fit

Before turning to forecasting, we first consider the ability of estimated factor models to summarize the data set. To this end, we extracted $q = 10$ factors using each of the three different loss functions. We selected the penalization parameter λ by minimizing the BIC (12); in all three cases, $\lambda = 0.1$ was selected. From Table 3

Table 3: Summary statistics for the in-sample fit in the macroeconomic data set.

	Nonzero	Approximation quality				Nonzero	Approximation quality		
Criterion	loadings	RMSE	MnAE	MdAE	Criterion	loadings	RMSE	MnAE	MdAE
L_2, $\lambda = 0$	1320	1.068	0.663	0.454	L_2, $\lambda = 0.1$	753	**1.061**	0.656	0.447
L_1, $\lambda = 0$	1320	1.246	**0.616**	**0.364**	L_1, $\lambda = 0.1$	842	1.258	0.622	0.365
Tukey, $\lambda = 0$	1320	1.081	0.626	0.422	Tukey, $\lambda = 0.1$	296	1.213	0.643	0.424

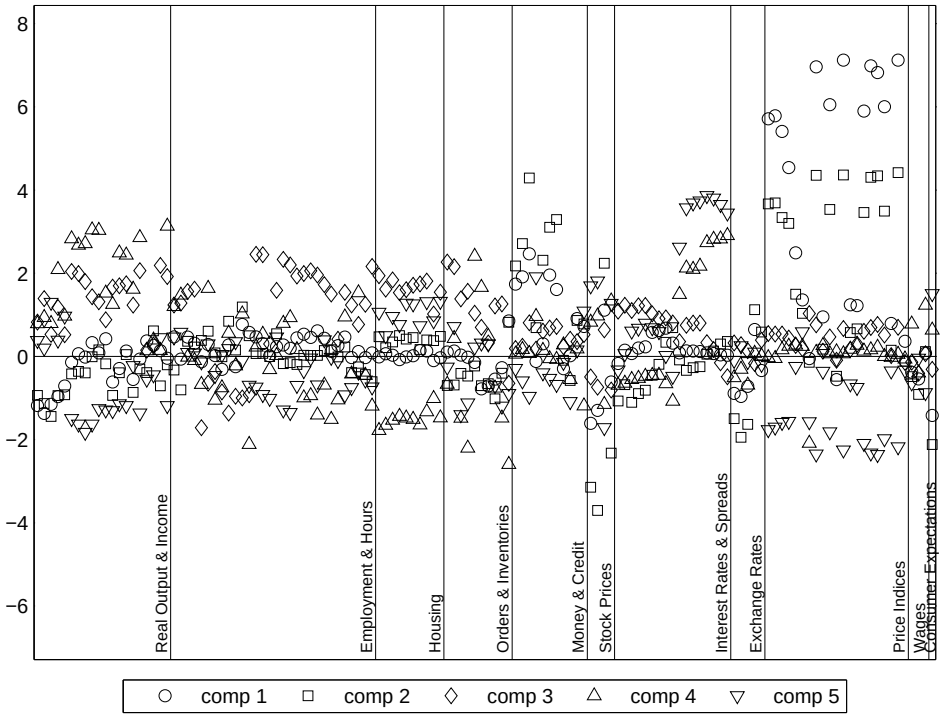
Notes: This table reports the number of nonzero entries in the estimated 132×10 loading matrix $\hat{A}$, as well as the root mean squared error and mean and median absolute error for the approximation $X \approx \hat{F}\hat{A}'$, after standardizing all variables to median zero and median absolute deviation one.

we note that, as expected, using the L_2 criterion leads to the smallest mean squared error $\left|\left|X - \hat{X}\right|\right|_2^2$, while using the L_1 criterion leads to the smallest mean and median absolute error. For all error measures, the Tukey criterion yields results in between these extremes. As for the simulated data in Section 3, we observe that setting a positive penalty term does not substantially influence the in-sample goodness of fit.

This table also clearly shows the sparsity effect of choosing $\lambda > 0$, leading to as few as 296 (out of 1320) nonzero factor loadings for the Tukey criterion. Figures 2 and 3 show how this property aids in the interpretation of the factors. In these figures, the variable number is on the horizontal axis, with groups of variables separated by vertical lines. The factor loading is on the vertical axis, and exact zero loadings were omitted for legibility. The factor loadings obtained by standard PCA (Figure 2) are quite difficult to interpret. (Stock and Watson (2002) resort to computing pairwise correlations between constructed factors and original variables to alleviate this problem.) On the other hand, Figure 3 allows for a reasonable interpretation of all ten factors extracted using the penalized Tukey criterion. For example, the pattern of nonzero loadings on the first component (circles in the top panel of Figure 3) suggests that this component is mostly associated with employment-related series. Continuing in this manner, we can assign labels to all ten factors as follows:

1. employment;
2. interest rates;
3. production;
4. interest rate spreads;
5. consumer price inflation;
6. housing;
7. producer price inflation;
8. exchange rates;
9. monetary policy; and
10. stock prices.

Figure 2: Nonzero factor loadings for the macroeconomic data, L_2 criterion, $\lambda = 0$.



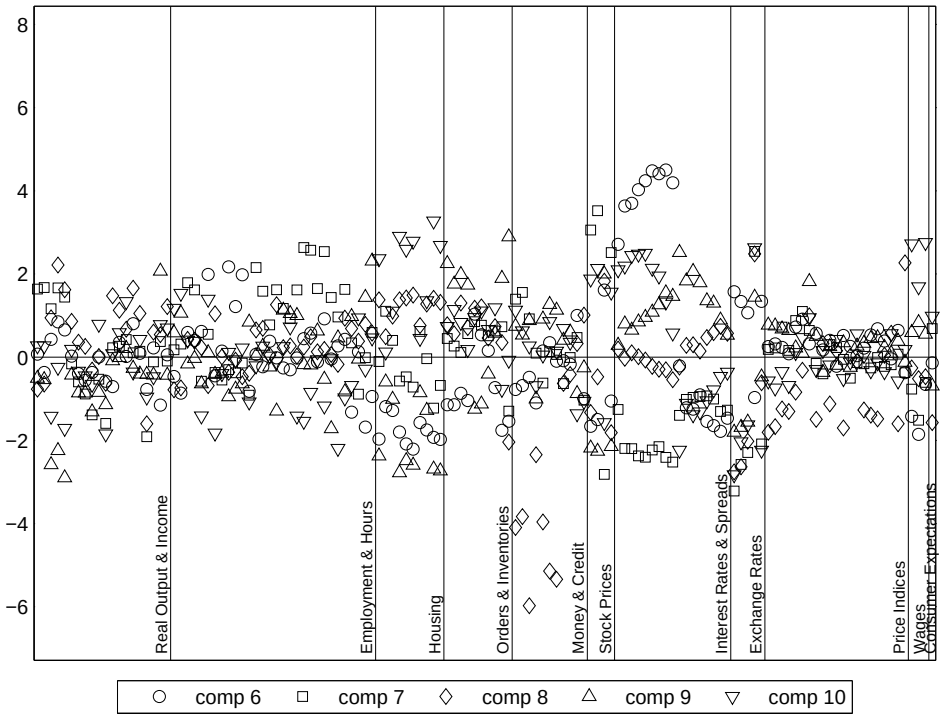


Figure 3: Nonzero factor loadings for the macroeconomic data, Tukey criterion, $\lambda = 0.1$.

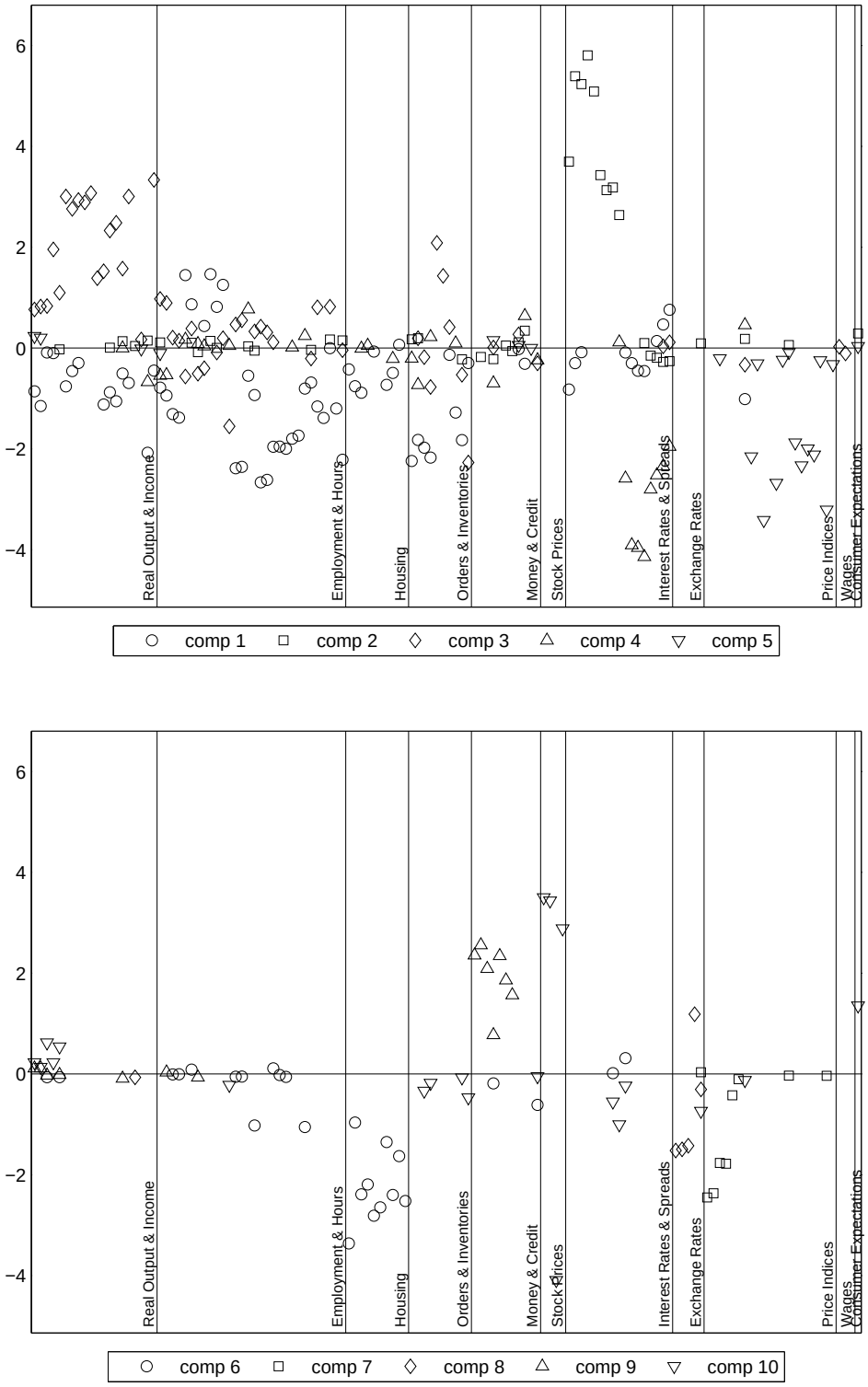
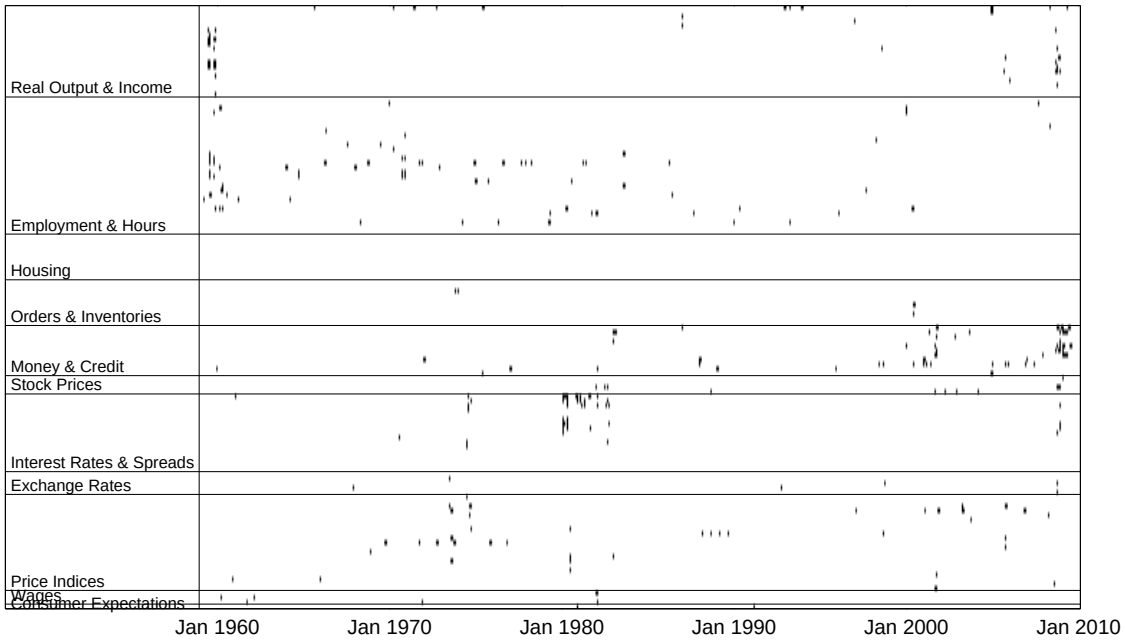


Figure 4: Heat maps of the residuals for the macroeconomic data. Top: L_2 criterion, $\lambda = 0$. Bottom: Tukey, $\lambda = 0.1$.



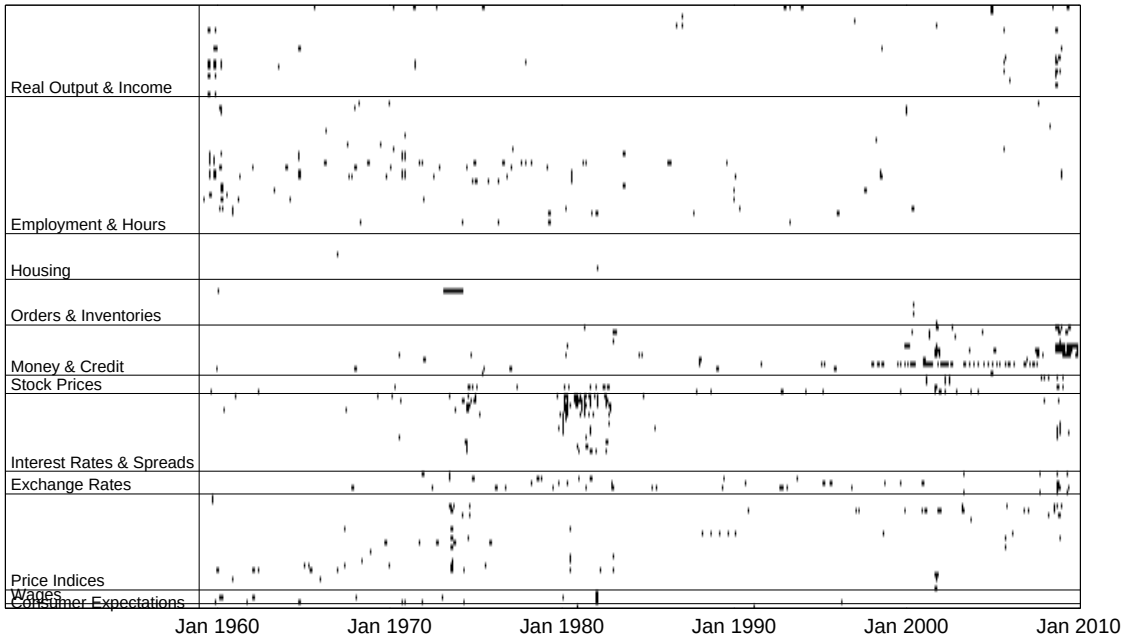


Table 4: Forecasting results for the macroeconomic data set: Industrial Production and Personal Income.

Horizon	Criterion	RMSE	MnAE	MdAE	Horizon	Criterion	RMSE	MnAE	MdAE
	Industrial Production					Personal Income			
$h = 1$	$L_2, \lambda = 0$	8.258	5.917	4.395	$h = 1$	$L_2, \lambda = 0$	5.723	3.703	2.716
	$L_2, \lambda > 0$	8.368	5.961	4.357		$L_2, \lambda > 0$	5.932	3.706	2.786
	$L_1, \lambda = 0$	**7.889**	**5.717**	**4.161**		$L_1, \lambda = 0$	5.416	3.550	2.628
	$L_1, \lambda > 0$	8.023	5.742	4.238		$L_1, \lambda > 0$	5.430	3.563	2.587
	Tukey, $\lambda = 0$	7.944	5.720	4.322		Tukey, $\lambda = 0$	**5.390**	**3.505**	2.642
	Tukey, $\lambda > 0$	7.969	5.768	4.422		Tukey, $\lambda > 0$	5.414	3.537	**2.563**
$h = 3$	$L_2, \lambda = 0$	5.811	4.352	3.350	$h = 3$	$L_2, \lambda = 0$	3.369	2.521	1.945
	$L_2, \lambda > 0$	5.834	4.347	3.338		$L_2, \lambda > 0$	3.387	2.539	2.038
	$L_1, \lambda = 0$	5.792	4.305	3.455		$L_1, \lambda = 0$	3.403	2.541	**1.923**
	$L_1, \lambda > 0$	**5.750**	**4.300**	3.347		$L_1, \lambda > 0$	**3.364**	**2.513**	1.981
	Tukey, $\lambda = 0$	5.927	4.346	**3.171**		Tukey, $\lambda = 0$	3.515	2.575	1.997
	Tukey, $\lambda > 0$	5.927	4.351	3.243		Tukey, $\lambda > 0$	3.415	2.547	2.101
$h = 6$	$L_2, \lambda = 0$	4.933	3.682	2.760	$h = 6$	$L_2, \lambda = 0$	**2.775**	2.141	1.689
	$L_2, \lambda > 0$	4.875	**3.617**	2.756		$L_2, \lambda > 0$	2.792	2.148	1.728
	$L_1, \lambda = 0$	**4.867**	3.758	3.080		$L_1, \lambda = 0$	2.880	2.100	1.598
	$L_1, \lambda > 0$	4.925	3.802	3.115		$L_1, \lambda > 0$	2.841	**2.081**	**1.545**
	Tukey, $\lambda = 0$	5.281	3.820	**2.672**		Tukey, $\lambda = 0$	3.025	2.209	1.625
	Tukey, $\lambda > 0$	4.965	3.684	2.673		Tukey, $\lambda > 0$	3.011	2.235	1.697
$h = 12$	$L_2, \lambda = 0$	3.825	**2.769**	2.051	$h = 12$	$L_2, \lambda = 0$	2.486	1.957	1.557
	$L_2, \lambda > 0$	**3.821**	2.775	2.165		$L_2, \lambda > 0$	**2.447**	1.937	1.557
	$L_1, \lambda = 0$	4.073	3.002	2.265		$L_1, \lambda = 0$	2.537	1.935	1.465
	$L_1, \lambda > 0$	3.996	2.947	2.243		$L_1, \lambda > 0$	2.487	**1.920**	**1.455**
	Tukey, $\lambda = 0$	4.001	2.862	**2.043**		Tukey, $\lambda = 0$	2.566	1.994	1.551
	Tukey, $\lambda > 0$	3.999	2.889	2.125		Tukey, $\lambda > 0$	2.534	1.951	1.546

Notes: This table reports the root mean squared forecast error and mean and median absolute forecast error for the macroeconomic forecasting example. For each series, the smallest RMSE, MeanAE, and MedianAE are printed in boldface.

Recalling the large number of outliers in the data, as visualized in the heat map in Figure 1, it is of interest to repeat this outlier detection exercise for the residuals after ten factors have been extracted. The corresponding heat maps are shown in Figure 4. We observe that the L_2 factor extraction procedure is severely influenced by the outlying observations identified in Figure 1: many of the outliers are no longer present in the residuals, which means that the extracted factors fit these observations well. This result continues to hold if a positive penalty λ is selected. On the other hand, the residuals from the Tukey criterion exhibit a similar outlier pattern to the original data. In this criterion, large outliers are downweighted, so that they have less impact on the factor estimates. Similar results are obtained using the L_1 criterion (not shown).

Table 5: Forecasting results for the macroeconomic data set: Manufacturing & Trade Sales and Employment.

Horizon	Criterion	RMSE	MnAE	MdAE	Horizon	Criterion	RMSE	MnAE	MdAE
	Manufacturing & Trade Sales					Employment			
$h = 1$	$L_2, \lambda = 0$	**11.463**	**8.680**	7.040	$h = 1$	$L_2, \lambda = 0$	**2.980**	2.227	**1.708**
	$L_2, \lambda > 0$	11.540	8.774	6.990		$L_2, \lambda > 0$	3.045	2.277	1.779
	$L_1, \lambda = 0$	11.779	8.963	7.246		$L_1, \lambda = 0$	2.991	**2.226**	1.710
	$L_1, \lambda > 0$	11.819	9.021	7.449		$L_1, \lambda > 0$	2.983	2.229	1.771
	Tukey, $\lambda = 0$	12.072	9.028	6.795		Tukey, $\lambda = 0$	3.072	2.307	1.778
	Tukey, $\lambda > 0$	12.108	9.066	**6.669**		Tukey, $\lambda > 0$	3.071	2.293	1.761
$h = 3$	$L_2, \lambda = 0$	6.205	4.689	3.648	$h = 3$	$L_2, \lambda = 0$	1.765	1.322	**0.984**
	$L_2, \lambda > 0$	6.363	4.781	3.747		$L_2, \lambda > 0$	1.773	1.336	1.025
	$L_1, \lambda = 0$	6.201	4.719	3.787		$L_1, \lambda = 0$	**1.733**	**1.296**	0.987
	$L_1, \lambda > 0$	**6.074**	**4.660**	3.705		$L_1, \lambda > 0$	1.757	1.323	1.015
	Tukey, $\lambda = 0$	6.297	4.763	**3.625**		Tukey, $\lambda = 0$	1.770	1.343	1.044
	Tukey, $\lambda > 0$	6.345	4.802	3.672		Tukey, $\lambda > 0$	1.780	1.338	1.038
$h = 6$	$L_2, \lambda = 0$	**4.663**	**3.406**	2.509	$h = 6$	$L_2, \lambda = 0$	**1.422**	**1.076**	**0.820**
	$L_2, \lambda > 0$	4.757	3.448	2.567		$L_2, \lambda > 0$	1.435	1.093	0.827
	$L_1, \lambda = 0$	5.127	3.695	2.605		$L_1, \lambda = 0$	1.456	1.108	0.837
	$L_1, \lambda > 0$	4.920	3.603	2.728		$L_1, \lambda > 0$	1.444	1.107	0.845
	Tukey, $\lambda = 0$	4.922	3.538	**2.367**		Tukey, $\lambda = 0$	1.524	1.143	0.823
	Tukey, $\lambda > 0$	4.868	3.494	2.467		Tukey, $\lambda > 0$	1.525	1.137	0.839
$h = 12$	$L_2, \lambda = 0$	3.664	2.613	**1.931**	$h = 12$	$L_2, \lambda = 0$	1.235	0.932	0.685
	$L_2, \lambda > 0$	**3.557**	**2.607**	2.016		$L_2, \lambda > 0$	**1.194**	**0.904**	**0.671**
	$L_1, \lambda = 0$	3.740	2.734	2.110		$L_1, \lambda = 0$	1.228	0.913	0.710
	$L_1, \lambda > 0$	3.714	2.731	2.156		$L_1, \lambda > 0$	1.238	0.914	0.711
	Tukey, $\lambda = 0$	3.630	2.679	2.121		Tukey, $\lambda = 0$	1.294	0.979	0.747
	Tukey, $\lambda > 0$	3.579	2.656	2.013		Tukey, $\lambda > 0$	1.290	0.978	0.737

4.3 Forecasting Results

Next, we focus on forecasting four key macroeconomic series, as described above. The results are summarized in Tables 4 and 5. For Industrial Production and Personal Income (Table 4), we find that our sparse and robust methods often outperform the benchmark of standard PCA. For horizons shorter than a year, the more robust Tukey and L_1 criterions generally lead to better forecasts than the standard L_2 criterion, irrespective of which measure we use to evaluate the performance. Thus, the lack of robustness in PCA that we observed negatively affects the forecasting performance, and more robust criterion functions remedy this situation.

For the easier task of forecasting annual growth rates ($h = 12$), we find that the L_2 criterion does lead to adequate forecasts. In this case, however, imposing a sparsity constraint improves the forecast quality: a relatively simple task is best performed using relatively simple models.

The results for the other two series, Manufacturing & Trade Sales and Employment, are shown in Table 5. It is very hard to improve on standard PCA forecasts for these series, a finding that was also documented by Exterkate et al. (2011). Nevertheless, our result that sparse modelling leads to better forecasts for annual growth rates also applies here.

5 Application: Boston Housing Data

5.1 Data and Forecast Model

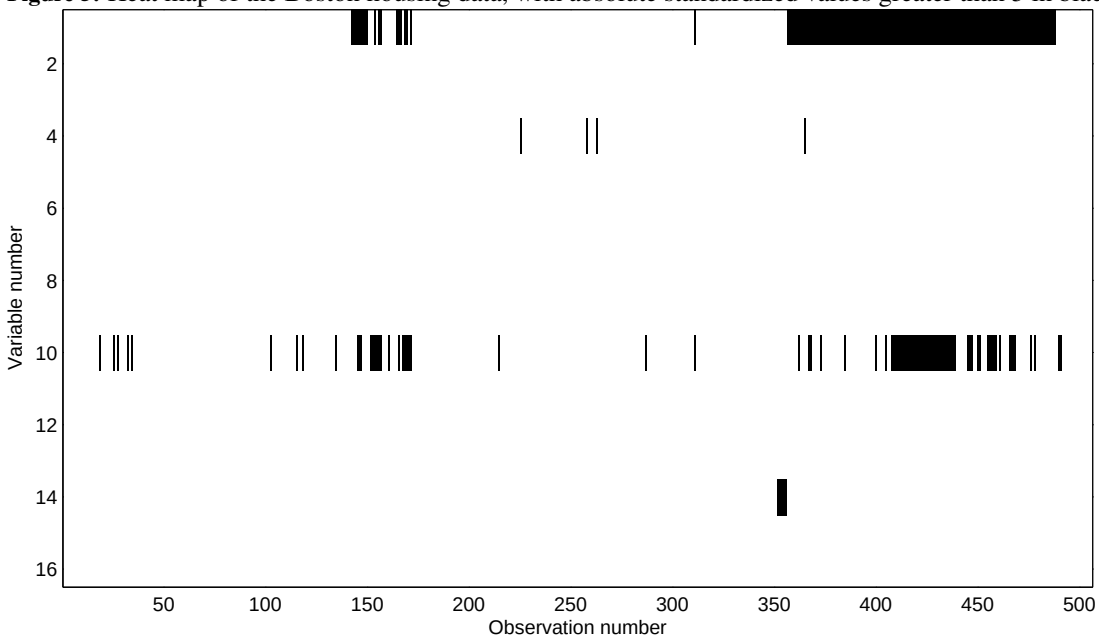
The Boston Housing data set, originating with Harrison and Rubinfeld (1978), has been extensively analyzed in the robust statistics literature. We use the corrected version of the data set by Pace and Gilley (1997). The data set contains various characteristics of houses, demographics, air pollution, and geographical details on 506 census tracts in or nearby Boston. The objective is to relate the median house price to the other characteristics, and our model will be inspired by the one in Pace and Gilley (1997):

$$\log \text{ Price} = \begin{pmatrix} 1, & \text{ZN}, \ \text{CHAS}, \ \text{CRIM}, \ \text{INDUS}, \ \text{NOX}^2, \ \text{RM}^2, \ \text{AGE}, \ \ldots \\ & \log \text{ DIS}, \ \log \text{ RAD}, \ \text{TAX}, \ \text{PTRATIO}, \ \text{B}, \ \log \text{ LSTAT}, \ \ldots \\ & \text{LON}, \ \text{LAT}, \ \text{LON} \times \text{LAT}, \ \text{LON}^2, \ \text{LAT}^2 \end{pmatrix} \beta + \varepsilon, \tag{16}$$

where the regressors denote the proportion of area zoned with large lots (ZN), a dummy for a location contiguous to the Charles River (CHAS), the crime rate (CRIM), the proportion of nonretail business areas (INDUS), levels of nitrogen oxides (NOX), the average number of rooms (RM), the proportion of structures built before 1940 (AGE), weighted distances to the employment centers (DIS), an index of accessibility (RAD), the property tax rate (TAX), the pupil/teachter ration (PTRATIO), the black population proportion (B), the lower status population proportion (LSTAT), and the geographical longitude (LON) and latitude (LAT).

Before applying the methods of Section 2 to this data set, we remove two variables for which the median absolute deviation is zero; namely, the proportion of large lots (ZN) and the Charles River dummy (CHAS). As we scale all variables by dividing by their median absolute deviations before extracting factors using the Tukey criterion, we cannot handle these variables in our algorithm. As a compromise, we estimate the model

Figure 5: Heat map of the Boston housing data, with absolute standardized values greater than 5 in black.



$$\log \text{ Price} = \alpha + \beta_1 \text{ ZN} + \beta_2 \text{ CHAS} + F\gamma + \varepsilon, \tag{17}$$

where the factors in F are extracted from the remaining right-hand-side variables in Equation (16).

A heat map of the data is shown in Figure 5, with the variables ordered as in Equation (16), starting with CRIM. Thus, the variables containing relatively many outlying observations can be identified as the crime rate (CRIM, variable 1) and the proportion of black population (B, variable 10). The groups of observations at which these outliers occur correspond to the locations in the cities of Cambridge (around observation 150) and Boston (around observations 400-450).

Our forecasting procedure is as follows. We first extract the factors F from the full data set. Then, we estimate Model (17) on a random selection of 80% of the 506 observations, and we decide on the number of factors, the value of λ, and whether or not to include ZN and/or CHAS in the model by minimizing a Bayesian Information Criterion similar to the one in Equation (15). The selected model is then used to forecast the prices for the 20% of the observations that were left out in the estimation, and we repeat the procedure five times, ensuring that each observation is being predicted exactly once.

Table 6: Summary statistics for the in-sample fit in the Boston housing data set.

	Approximation quality				Approximation quality		
Criterion	RMSE	MnAE	MdAE	Criterion	RMSE	MnAE	MdAE
L_2, $\lambda = 0$, $q = 5$	4.683	1.159	0.186	L_2, $\lambda = 0.001$, $q = 5$	4.687	1.153	0.185
L_1, $\lambda = 0$, $q = 5$	6.219	0.915	**0.039**	L_1, $\lambda = 0.100$, $q = 5$	6.600	0.985	0.056
Tukey, $\lambda = 0$, $q = 5$	**1.619**	**0.391**	0.204	Tukey, $\lambda = 0.010$, $q = 5$	5.496	0.748	0.172

Notes: This table reports the selected numbers of factors and penalization parameters, as well as the root mean squared error and mean and median absolute error, after standardizing all variables to median zero and median absolute deviation one. The smallest errors are printed in boldface.

Table 7: Forecasting results for the Boston housing data set.

Criterion	RMSE	MnAE	MdAE	Criterion	RMSE	MnAE	MdAE
L_2, $\lambda = 0$	0.224	0.148	0.100	L_2, $\lambda > 0$	**0.217**	**0.142**	**0.097**
L_1, $\lambda = 0$	0.240	0.156	0.097	L_1, $\lambda > 0$	0.241	0.157	0.100
Tukey, $\lambda = 0$	0.269	0.191	0.130	Tukey, $\lambda > 0$	0.233	0.152	0.100

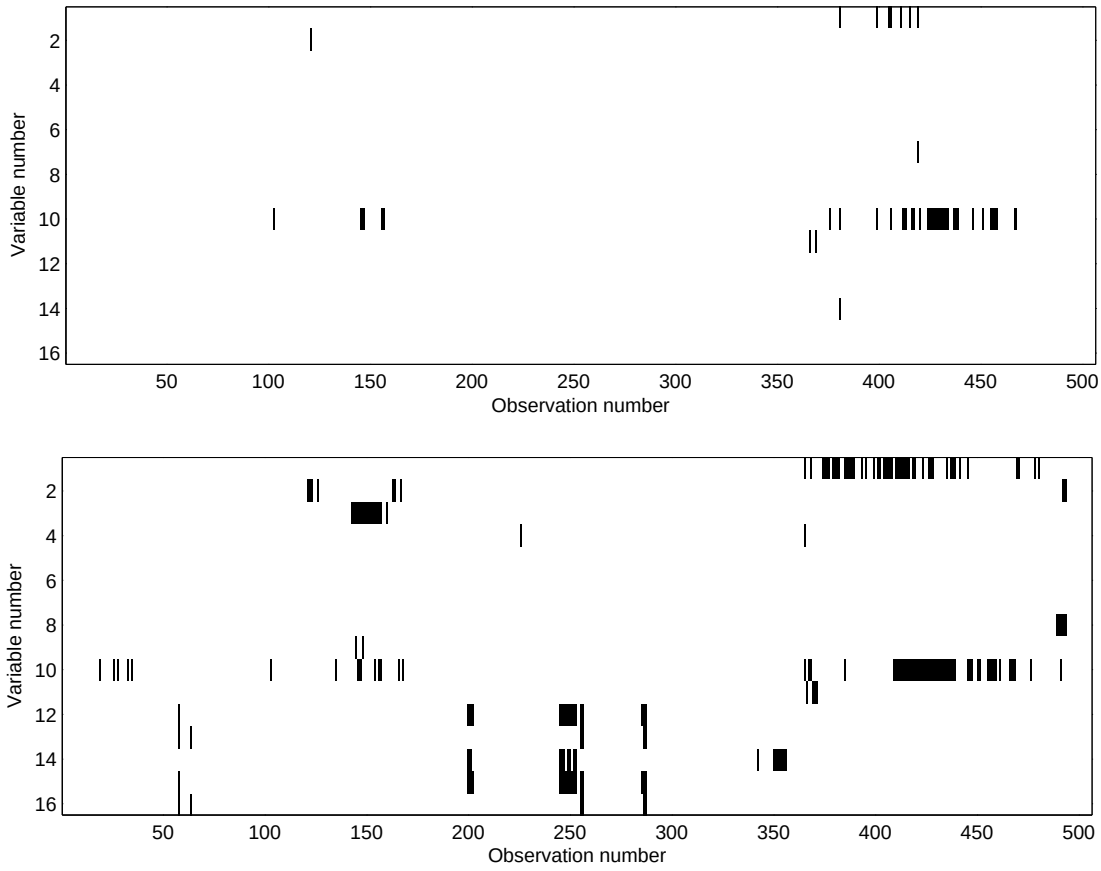
Notes: This table reports the root mean squared error and mean and median absolute error in forecasting the logarithm of the median house price. The smallest errors are printed in boldface.

5.2 In-Sample Fit

Again, we first consider the in-sample fit, selecting the number of factors and the penalization parameter by minimizing the BIC in Equation (12). Summary statistics are shown in Table 6. Note that in all cases, the maximum number of five components is selected. Given that the data set contains only sixteen variables, it seems undesirable to extract even more factors. As in the other data sets, we find that allowing for a positive penalization parameter λ does not substantially worsen the in-sample fit — except, in this case, for the Tukey criterion, which performs extremely well with $\lambda = 0$.

Heat maps of the residuals are shown in Figure 6. The heat map for standard PCA residuals (top panel) indicates that many of the outlying observations that were identified from Figure 5 are fitted by the factor structure. This is the well-known effect of least-squares methods being sensitive to large outliers. On the other hand, most of the outliers are still present in the residuals from penalized Tukey factor extraction. In fact, new groups of outliers are now detected, especially in the variables numbered 12 (geographical longitude), 14 (longitude times latitude), and 15 (longitude squared). It turns out that the outliers correspond to locations relatively far to the west of Boston. The Tukey criterion tries to fit *most* of the data, rather than *all* of the data.

Figure 6: Heat maps of the residuals for the Boston housing data. Top: L_2 criterion, $\lambda = 0$. Bottom: Tukey, $\lambda = 0.010$.

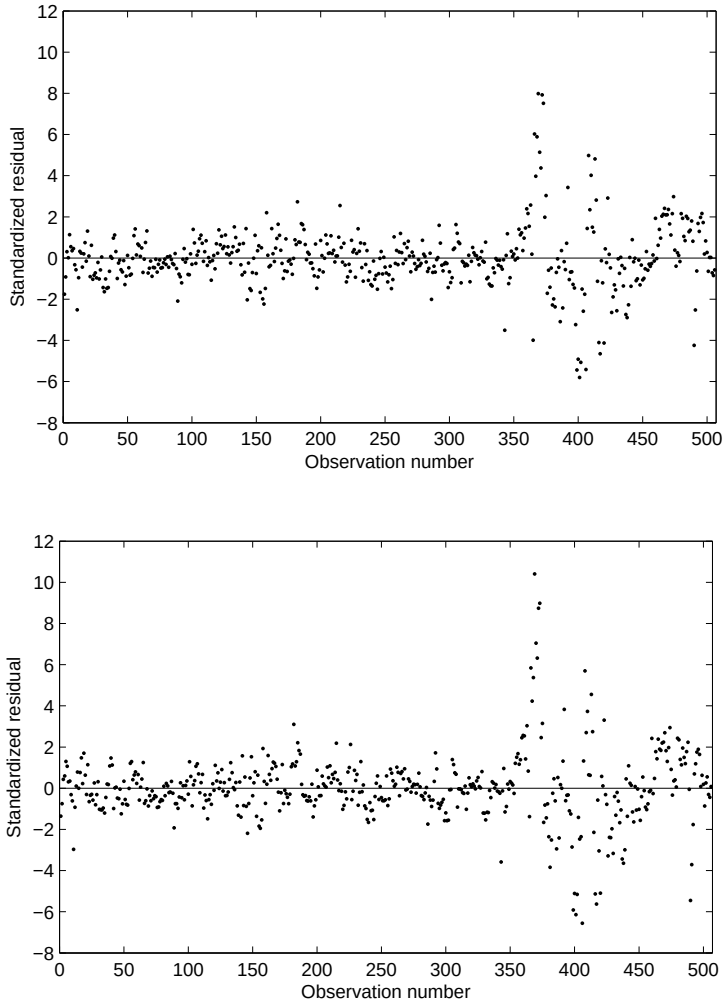


5.3 Forecasting Results

Table 7 summarizes the forecasting results. We observe that even in this relatively small data set, penalized estimation leads to better forecasts. On the other hand, despite the large number of outliers identified in Figure 5, robust methods perform (somewhat) worse than L_2 estimation. Closer inspection of the data reveals that the house price has a high correlation with the two variables containing most outliers. Thus, in this case, it is undesirable to downweight the outlying observations, but the results are not too heavily affected by this fact.

To illustrate this result, we re-estimated Equation (17) over the full sample, again selecting the number of factors and the value of λ by minimizing the BIC. The residuals, standardized to median absolute deviation one, are plotted in Figure 7 for the L_2 and Tukey factor estimates. These plots show that although both methods have difficulties fitting the house prices for the city of Boston (around observation 400), using the

Figure 7: Standardized residuals for the house price equation (17). Top: L_2 criterion, $\lambda = 0.1$, $q = 5$. Bottom: Tukey criterion, $\lambda = 0.1$, $q = 4$.



more robust Tukey factors leads to greater errors for these observations than using L_2-based factors. This result implies that these outliers can be considered "good leverage points": downweighting them adversely affects the fit of the model.

6 Conclusion

We propose a novel factor extraction method that unifies two recent strands in the factor modelling literature, robustness and sparsity. This method leads to a sparse factor loading matrix and to factors that are robust to

outlying observations in the original data. We are the first to combine these two issues in the context of factor modelling, and we argue that both properties can be helpful in forecasting. A Monte Carlo study confirms this intuition: compared to standard principal component analysis, our proposed method gives a much closer approximation to the true factor space; hence, it is more suitable for forecasting purposes. This improvement is obtained at the cost of only small losses in in-sample fit.

We apply this method to two economic data sets. Our first application concerns macroeconomic forecasting using a large panel of predictors. We show that, compared to traditional principal component analysis, our proposed method leads to more interpretable factors. Moreover, we report favorable forecasting performance: for annual growth rates, imposing sparsity on the factor loadings leads to more accurate forecasts for all target variables considered. For shorter-term growth rates, robust estimation provides an additional advantage in forecasting U.S. Industrial Production and Personal Income. This result shows that our factor extraction method, which can be thought of as "multivariate data cleaning", is useful even after the standard univariate data cleaning that was performed by Stock and Watson (2002).

In the second economic application, we analyze the well-known Boston Housing data set. Even in this relatively small data set (sixteen predictor variables), we find that sparse estimation improves the quality of forecasts. We also argue that robust techniques can be expected to fare worse in this data set, as the outliers are actually "good leverage points"; however, their impact on the forecast accuracy turns out to be minimal.

We note that if prior knowledge on a sparse factor structure is available, it is of course possible to impose that certain elements of the loading matrix are zero and use more traditional factor extraction methods. This is the case in the macroeconomic data set analyzed in Section 4 of this paper, in which the series are categorized into groups of related variables. However, the results in Section 4.2 show that even in this case, the selection of factor loadings that our methodology sets to zero in a data-driven way is similar to the selection that we would impose to be zero based on prior information.

To conclude, we find that sparse and robust estimation of factor models has a great potential for improving both the interpretability of the estimated factors and the accuracy of forecasts. Given its favorable performance in a macroeconomic forecasting study, an interesting generalization of our method would be to dynamic factor models, in which explicit assumptions about the evolution of the factors over time are made.

References

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146: 304–317, 2008.

M. Bańbura, D. Giannone, and L. Reichlin. Large Bayesian vector autoregressions. *Journal of Applied Econometrics*, 25:71–92, 2010.

A. Carriero, G. Kapetanios, and M. Marcellino. Forecasting large datasets with Bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, 26:in press, 2011.

C. Croux, P. Filzmoser, G. Pison, and P.J. Rousseeuw. Fitting multiplicative models by robust alternating regressions. *Statistics and Computing*, 13:23–36, 2003.

F. De la Torre and M.J. Black. Robust principal component analysis for computer vision. In *International Conference on Computer Vision*, pages 362–369, Vancouver, Canada, 2001.

P. Exterkate, P.J.F. Groenen, C. Heij, and D. van Dijk. Nonlinear forecasting with many predictors using kernel ridge regression. *Tinbergen Institute Discussion Paper No. 11-007*, 2011.

G. Fagiolo, M. Napoletano, and A. Roventini. Are output growth-rate distributions fat-tailed? Some evidence from OECD countries. *Journal of Applied Econometrics*, 23:639–669, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

D. Harrison and D.L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

I.T. Jolliffe, N.T. Trendafilov, and M. Uddin. A modified principal component technique based on the Lasso. *Journal of Computational and Graphical Statistics*, 12:531–547, 2003.

S.C. Ludvigson and S. Ng. The empirical risk-return relation: A factor analysis approach. *Journal of Financial Economics*, 83:171–222, 2007.

S.C. Ludvigson and S. Ng. Macro factors in bond risk premia. *Review of Financial Studies*, 22:5027–5067, 2009.

R.A. Maronna and V.J. Yohai. Robust low-rank approximation of data matrices with elementwise contamination. *Technometrics*, 50:295–304, 2008.

R.A. Maronna, D.R. Martin, and V.J. Yohai. *Robust statistics: Theory and methods*. Wiley, New York, 2006.

R.K. Pace and O.W. Gilley. Using the spatial configuration of the data to improve estimation. *Journal of Real Estate Finance and Economics*, 14:333–340, 1997.

G. Pison, P.J. Rousseeuw, P. Filzmoser, and C. Croux. Robust factor analysis. *Journal of Multivariate Analysis*, 84:145–172, 2003.

J.H. Stock and M.W. Watson. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics*, 20:147–162, 2002.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58:267–288, 1996.

H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business and Economic Statistics*, 25:347–355, 2007.

D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal component analysis and canonical correlation analysis. *Biostatistics*, 10:515–534, 2009.

H. Wold. Nonlinear estimation by iterative least squares procedures. In F. David, editor, *Research papers in statistics: Festschrift for J. Neyman*, pages 411–444. Wiley, New York, 1966.

H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the Lasso. *Annals of Statistics*, 35: 2173–2192, 2007.