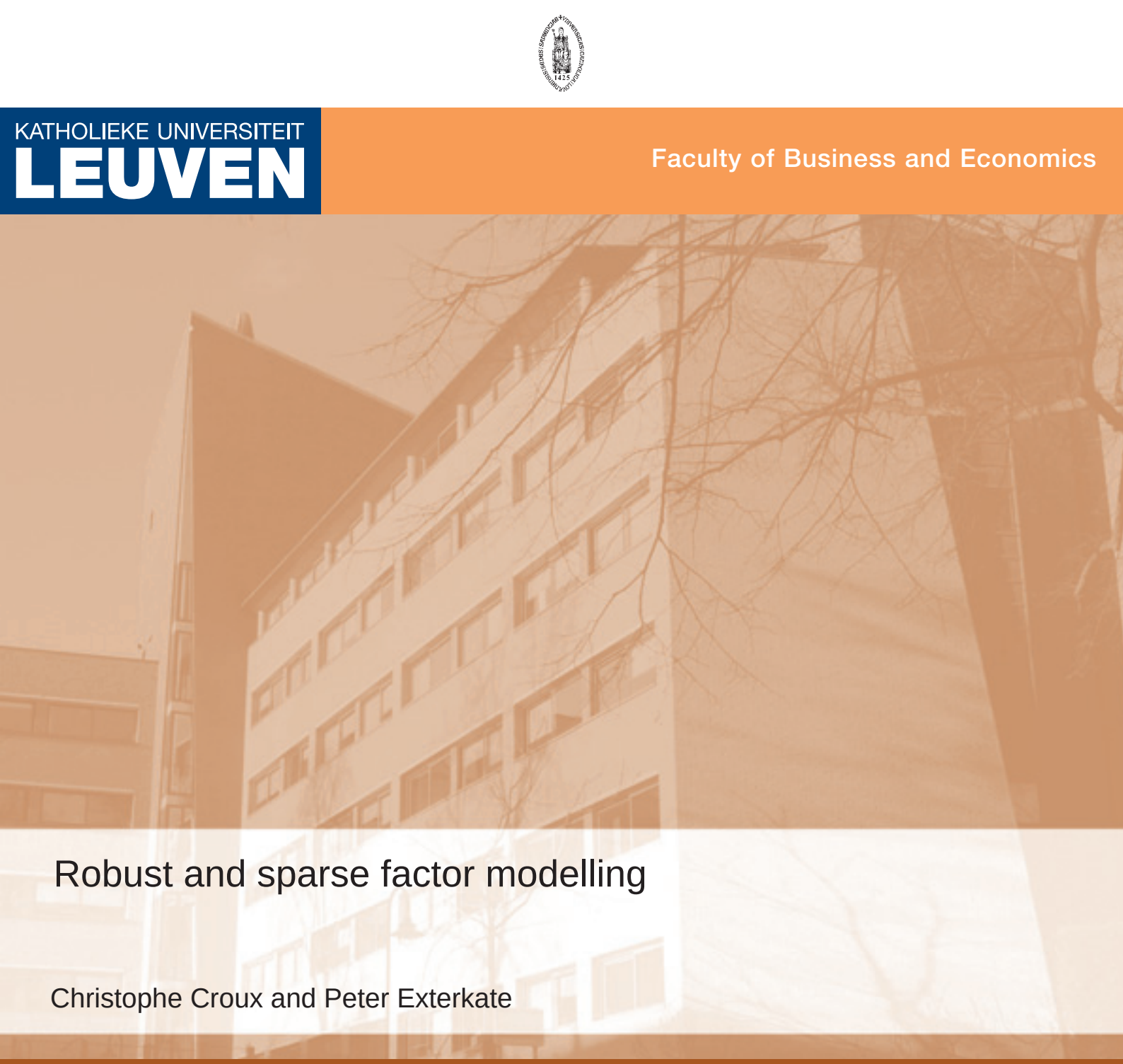

KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Robust and sparse factor modelling

Christophe Croux and Peter Exterkate

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1120

Robust and Sparse Factor Modelling

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

Peter Exterkate
Erasmus School of Economics
Erasmus University Rotterdam

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 the traditional factor construction method, we find that this procedure leads to a favorable forecasting performance in the presence of outliers and to better interpretable factors. We investigate the performance of the method in a Monte Carlo experiment and in an empirical application to a large data set from macroeconomics.

Keywords: dimension reduction, forecasting, outliers, regularization, sparsity.
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 (Croux and Haesbroeck, 2000; 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 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 an empirical application. The simulation results show that ignoring the presence of outlying observations, which are often overlooked in empirical econometric studies, has important consequences for forecast accuracy. The application concerns forecasting key U.S. macroeconomic variables, as in Stock and Watson (2002).

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, little attention has been given to robustness issues in this context. Outlier-resistant estimators have typically only been applied to econometric models with a smaller number of variables (e.g. Fagiolo et al., 2008; Dehon et al., 2009). 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. An empirical application to macroeconomic forecasting follows in Section 4, and Section 5 concludes.

2 Methodology

2.1 Robust Data Matrix Approximation

We consider the problem of approximating an $n \times p$ data 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 least squares regression approach (due to Wold, 1966) is also possible. Given an initial estimates for F, we iterate until convergence:

- For a given F, minimize (1) with respect to A by solving p ordinary least squares (OLS) problems: the jth row of A is $a_j = (F'F)^{-1} F'x_j$, where x_j denotes the jth column of X.
- For a given A, minimize (1) with respect to F by solving n OLS problems: the ith row of F 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 (see e.g. 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), among others, who labelled it Alternating L_1 Regressions.

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

$$\rho(r) = \min \left\{ 1, \left(1 - (r/c)^2 \right)^3 \right\}. \tag{3}$$

This loss function is bounded, which makes it very robust to large outliers. The constant c is fixed at 3.4437, so that an 85% statistical 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 n residuals $x_{ij} - f_i' a_j$. Then, Maronna and Yohai (2008) propose to minimize

$$Q_{\text{Tukey}}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \hat{\sigma}_j^2 \sum_{i=1}^{n} \rho \left(\frac{x_{ij} - f_i' a_j}{\hat{\sigma}_j} \right). \tag{4}$$

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

$$\hat{\sigma}_j = 1.48 \underset{i}{\text{median}} \left\{ \left| x_{ij} - f_i' a_j \right| \right\}, \tag{5}$$

where the factor 1.48 ensures consistent scale estimation at normal distributions.

If we would set $\rho(r) = r^2$, criterion (4) reduces to the classical PCA criterion (1). In order to be able to apply the alternating algorithm to minimize (4), we rewrite it as an iteratively reweighted least squares problem. Defining weights

$$w_{ij} = \left(\frac{x_{ij} - f_i' a_j}{\hat{\sigma}_j} \right)^{-2} \rho \left(\frac{x_{ij} - f_i' a_j}{\hat{\sigma}_j} \right), \tag{6}$$

the objective in equation (4) can be rewritten as

$$Q_{\text{Tukey}}(F, A; X) = \frac{1}{2n} \sum_{j=1}^{p} \sum_{i=1}^{n} w_{ij} \left(x_{ij} - f_i' a_j\right)^2 . \tag{7}$$

This means that, given initial estimates of F and of the weights, we can solve (4) by iterating the following scheme until convergence:

- For a given F and given weights, minimize (7) with respect to A by solving p weighted least squares (WLS) problems: the jth row is $a_j = (F' D_j F)^{-1} F' D_j x_j$, 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 (5) and compute all weights w_{ij} using (6).
- For a given A and given weights, minimize (7) with respect to F by solving n WLS problems: the ith row is $f_i = (A' D_i A)^{-1} A' D_i x_i$, 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 (4), to median zero and median absolute deviation one. Initial estimates for F and the weights are obtained as described in 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 (4): for some positive scalar λ, called the *penalty parameter*, we aim to

minimize

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

where Q denotes either Q_{L_2}, Q_{L_1}, or Q_{Tukey}. As it stands, objective (8) 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 (8). As our purpose is not to impose sparsity on F, this additional term will be an L_2 penalty: we minimize

$$Q\left(F, A; X\right) + \lambda \sum_{j=1}^{p} \sum_{k=1}^{q} \left|a_{jk}\right| + \nu \sum_{i=1}^{n} \sum_{k=1}^{q} f_{ik}^2. \tag{9}$$

Finally, we note that Problem (9) is overparameterized: if the factorization $\left(\hat{F}, \hat{A}\right)$ solves (9) 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 (9). First, given F and (in the Tukey case) the weights w_{ij}, finding the jth row of A amounts to minimizing

$$Q\left(F, A; X\right) + \lambda \sum_{k=1}^{q} \left|a_{jk}\right|. \tag{10}$$

For the L_2 criterion function, we recognize (10) as a Lasso problem (Tibshirani, 1996). For the robust Tukey criterion, we need to perform a weighted Lasso regression, i.e. a Lasso with dependent variable $\left(\sqrt{w_{ij}}\right) x_{ij}$ and regressors $\left(\sqrt{w_{ij}}\right) f_i$, instead of a weighted least squares regression. Efficient algorithms to compute the Lasso solution are known; see Friedman et al. (2010). For the L_1 criterion, minimizing (10) 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

$$Q\left(F, A; X\right) + \frac{1}{2n} \sum_{k=1}^{q} f_{ik}^2. \tag{11}$$

For the L_2 and Tukey criteria, (11) is a ridge regression problem and can be solved analytically, resulting in

$$f_i = \left(A' D_i A + I_q\right)^{-1} A' D_i x_i. \tag{12}$$

Here I_q in the identity matrix of size q, and D_i is a diagonal matrix containing the weights $w_{i1}, \ldots, w_{ip}$. In the L_2 case, we have $D_i = I_p$. 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

$$BIC_{q,\lambda} = 2 \sum_{j=1}^{p} \log \hat{\sigma}_{j;q,\lambda} + \mathrm{df}_{q,\lambda} \cdot \frac{\log n}{n}. \tag{13}$$

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 loadings matrix A. Further, we approximate the determinant of the residual covariance matrix by the product of estimates of the p residual variances. 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. The scale estimate $\hat{\sigma}_{j;q,\lambda}$ is given by (5) when using the Q_{Tukey} criterion, by the mean absolute deviation when using Q_{L_1}, and by the standard deviation when using Q_{L_2}.

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 is $q = 2$. 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:

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

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 last two DGPs. If an observation is a vertical outlier, the latent factors behave normally but the observed variable is contaminated. For a bad leverage row both the factor variables and the noise term are outlying. The bad leverage rows are such that observed variables do not show any outlying value, making it difficult to detect them. Bad leverage points are considered to be the most dangerous, as is well documented in regression analysis (e.g. Verardi and Croux, 2009).

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)-(4) and the penalized criterion (9). In the latter case, the penalty parameter λ is selected by minimizing the BIC given in (13) over the grid $\log_{10} \lambda \in \{-4, -3, -2, -1, 0\}$. We treat the true number of factors ($q = 2$) as known in this simulation, to keep the computation time within limits.

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

		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 four 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.

Sparsity: 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 (where $\lambda = 0$) cannot succeed in exactly estimating zero loadings. The results for the penalized methods, on the other hand, are better. The penalized L_1 criterion performs best in identifying 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.

Forecast performance: 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 an error term. 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. 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 the columns of F and $\hat{F}$, respectively; the smaller this angle is, the more suitable $\hat{F}$ is for forecasting variables of the form $F\beta$.

Table 2: Simulated average angle between estimated and true factor space.

DGP	Criterion	Angle	DGP	Criterion	Angle
Normal	L_2, $\lambda = 0$	0.225	Vertical outliers	L_2, $\lambda = 0$	1.314
	L_2, $\lambda > 0$	**0.219**		L_2, $\lambda > 0$	1.332
	L_1, $\lambda = 0$	0.259		L_1, $\lambda = 0$	**0.286**
	L_1, $\lambda > 0$	0.256		L_1, $\lambda > 0$	0.288
	Tukey, $\lambda = 0$	0.233		Tukey, $\lambda = 0$	0.300
	Tukey, $\lambda > 0$	0.228		Tukey, $\lambda > 0$	0.291
Heavy tails	L_2, $\lambda = 0$	0.435	Bad leverage rows	L_2, $\lambda = 0$	1.264
	L_2, $\lambda > 0$	0.412		L_2, $\lambda > 0$	1.289
	L_1, $\lambda = 0$	0.295		L_1, $\lambda = 0$	0.344
	L_1, $\lambda > 0$	**0.291**		L_1, $\lambda > 0$	0.388
	Tukey, $\lambda = 0$	0.326		Tukey, $\lambda = 0$	0.325
	Tukey, $\lambda > 0$	0.311		Tukey, $\lambda > 0$	**0.320**

Notes: This table reports average results over 1000 replications of each of the four data-generating processes described in the text. We report the angle between the linear subspaces spanned by the columns of F and $\hat{F}$, in radians. For each DGP, the smallest angle is printed in boldface.

The average values of this angle, again over 1000 simulation runs, are reported in Table 2. Let us start comparing the unpenalized estimators ($\lambda = 0$). For the normal DGP the L_2 approach is the best, as expected. But the loss in precision by using the Tukey or L_1 approach remains limited. Under heavy tails the L_2 approach loses its optimality, and it gives the worst performance of all considered estimators. It becomes even more dramatic when outliers, either vertical or bad leverage rows, are present in the data. Then the L_2 approach, so using standard PCA, gives completely unreliable results, with an average angle close to $\pi/2 \approx 1.571$. This means that in the presence of outliers the factor space estimated by standard PCA is almost orthogonal to the true factor space, clearly showing its lack of robustness. The Tukey and L_1 approach continue to perform well, also in presence of outliers. In particular the Tukey criterion performs remarkably well in the case of bad leverage rows.

Let us now study the impact of adding a penalty parameter in the objective function. We first study the results for the DGP with normal or heavy tailed errors, reported in Table 2. We see that for this simulation design where the true factor structure is rather sparse, the sparse estimators improve on the unpenalized ones. This happens for the three criteria we considered. The gain in efficiency remains rather limited, however, and one would need an even stronger sparse structure of the true factors to make the advantage of the penalization

become more apparent. For the settings with outliers, either vertical or bad leverage rows, adding the penalty term only improves the performance of the most robust procedure, based on the Tukey criterion.

To summarize, we can state that both the L_1 and the Tukey criterion give good results, and outperform the standard L_2 approach by large margins if we deviate from the normal model. The gains in performance are mainly coming from the use of the robust loss functions, since adding the sparsity penalty term only slightly increases estimation precision further. One should not forget, however, that sparse solutions have the advantage of an easier interpretability of the loadings matrix.

4 Application: Macroeconomic Forecasting

4.1 Data and Forecasting Model

To evaluate the forecast performance of robustly and sparsely estimated factor models in an empirical application, we consider forecasting 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). 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. Stock and Watson (2002) define a partitioning of the data set into 11 economically meaningful groups of related variables.

We focus on forecasting four key measures of real economic activity: Industrial Production, Personal Income, Manufacturing & Trade Sales, and Employment. For each of these variables, we produce out-of-sample forecasts for the annualized h-month percentage growth rate, which is 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 $h = 1, 3, 6$ months.

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 (14) 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 (14), 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 (9) 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 (14) using either OLS, L_1 regression, or Tukey regression[1], with the same loss functions as used to extract the factors. Indeed, if there is a risk that outliers are present in the data, the forecast equation (14) needs to be estimated using robust regression. As the number of parameters in (14) 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 (13) for the BIC, which measures how well the factors $\hat{F}$ fit X. Instead, we minimize

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

[1]Tukey regression is better known as S-estimation of regression, and minimizes a robust scale estimator of the residuals, instead of the sum of squared residuals as for OLS. This robust scale estimator is defined using the Tukey biweight loss function in (3). S-estimators are resistant to vertical outliers and leverage points. See Maronna et al. (2006) for a complete description of S-estimators.

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

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

Notes: This table reports 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, together with the number of nonzero entries in the estimated 132×10 loading matrix $\hat{A}$.

where $(1 + \ell_y + \ell_f \cdot q)$ is the number of parameters in Equation (14), 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 2.3, 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.

4.2 In-Sample Fit

Before turning to forecasting, we first consider the ability of estimated factor models to summarize the data set. We extracted $q = 10$ factors using each of the three different loss functions. We selected the penalization parameter λ by minimizing the BIC (13), and estimate the factor and loading matrix. The residual matrix is then given by $X - \hat{F}\hat{A}'$. Table 3 summarizes the quality of the fit by computing the root mean squared error (RMSE), the mean absolute error (MnAE), and the median absolute error (MdAE) from the $n \times p$ residuals.

The L_2 approach gives the best in-sample RMSE, by construction. But if the quality of the fit is measured by other criteria, other methods do better. For the mean and median absolute error, the L_1 method gives the best results. For all considered goodness of fit measures, the Tukey method yields results between the L_1 and the L_2 approach.

Another interesting observation from Table 3 is that setting a positive penalty term does not substantially change the approximation quality. The gained sparsity comes at almost no loss in goodness of fit. The table

also reports the number of estimated nonzero loadings. We see that the sparse methods deliver a substantial number of zero estimated loadings, with an almost negligible effect on the quality of the in-sample fit.

Finally, note that the RMSE for the unpenalized L_2 criterion is slightly larger than for $\lambda > 0$, with a difference in the fourth significant digit. Mathematically, this is not possible, since the unconstrained L_2 method minimizes RMSE. The observed difference is due to numerical approximation error, the estimates being computed with the iterative alternating regression algorithm described in Section 2.1. Since the data set contains missing values, the L_2 method could not be computed using a standard singular value decomposition. The alternating regression scheme, however, can cope with missing data.

4.3 Robustness and Sparsity

In this section we focus on the estimates obtained by the Tukey criterion with penalization. Similar results are obtained using the L_1 criterion (not shown). We show two types of graphical displays useful for (i) outlier detection (ii) factor interpretation. The first display requires robustness of the method, the second one requires sparsity.

Outlier detection: An outlier is an observation that is unlikely to follow the factor model. A large value of the residual indicates a potential outlier. As an outlier detection tool we propose to make a heat map of the standardized residuals $(x_{ij} - \hat{f}_i'\hat{a}_j)/\hat{\sigma}_j$, with $1 \leq i \leq 120$ and $1 \leq j \leq 132$. The heat map is shown in Figure 1, with on the horizontal axis the time index i, and on the vertical axis the variable index j. The grouping of the variables in the 11 categories given by Stock and Watson (2002) is indicated as well. If an entry of the matrix of standardized residuals is larger than five in absolute value, it is indicated in black on the heat map, flagging the outlier. It is crucial to diagnose outliers starting from a robust fit. Otherwise, the present outliers may substantially affect the (non-robust) estimates of factors and loadings, potentially resulting in outliers with small residuals (*masking effect*) or good observations with large residuals (*swamping effect*). The masking and swamping effect are avoided when the residuals are computed form a robust fit.

From the heat map in Figure 1 one sees that a relatively large number of outliers shows up. One discovers outliers 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.

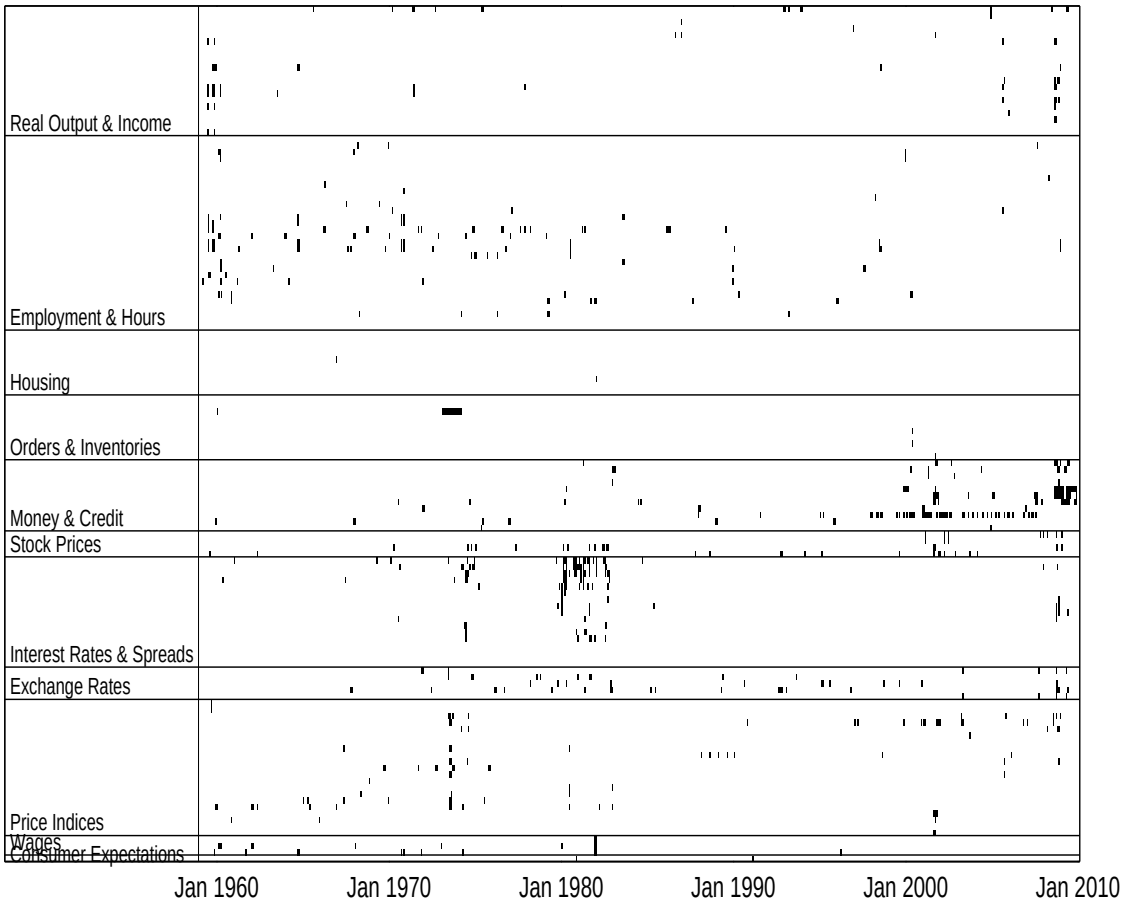

Figure 1: Heat map of the standardized residuals for the factor model using the Tukey criterion with $\lambda = 0.1$. Outliers are indicated in black.

Factor interpretation: Table 3 shows the sparsity effect of choosing $\lambda > 0$, leading to as few as 296 (out of 1320) nonzero factor loadings for the Tukey criterion. Figure 2 shows how sparsity aids in the interpretation of the factors. In this figure, the variable number is on the horizontal axis, with the 11 groups of variables separated by vertical lines. The factor loadings are on the vertical axis; the top panel contains the values of the loadings on the first 5 factors, each of them indicated by a different symbol, and the lower panel pictures the loadings on the last 5 factors. If the value of a loading is zero symbols are omitted, improving the legibility of the figure. A similar figure is made for a non-sparse method, see Figure 3, but this plot is much more difficult to interpret; we have more than four times as many symbols.

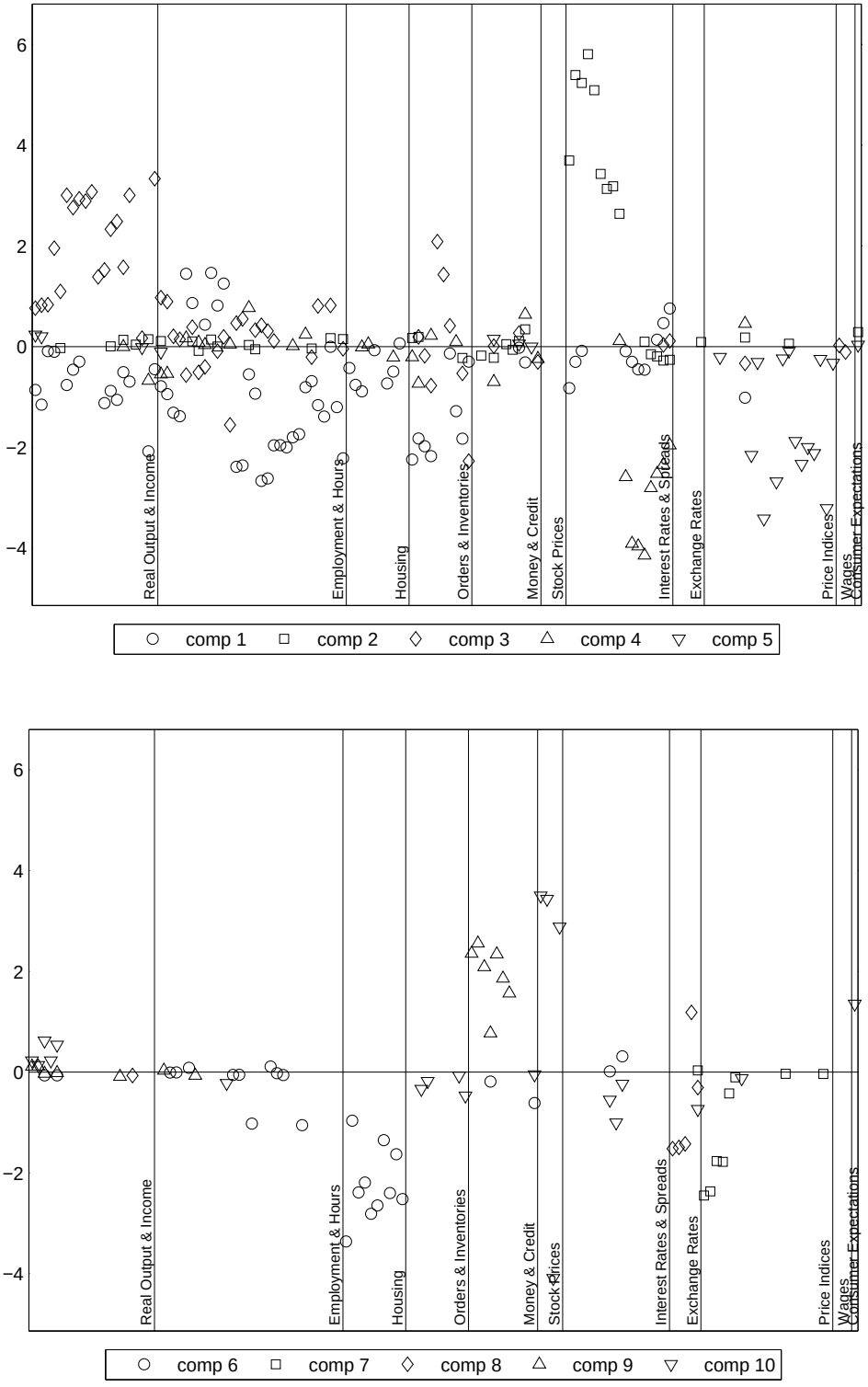

Figure 2: Nonzero factor loadings for the variables in the macroeconomic data set, using the Tukey criterion with $\lambda = 0.1$.

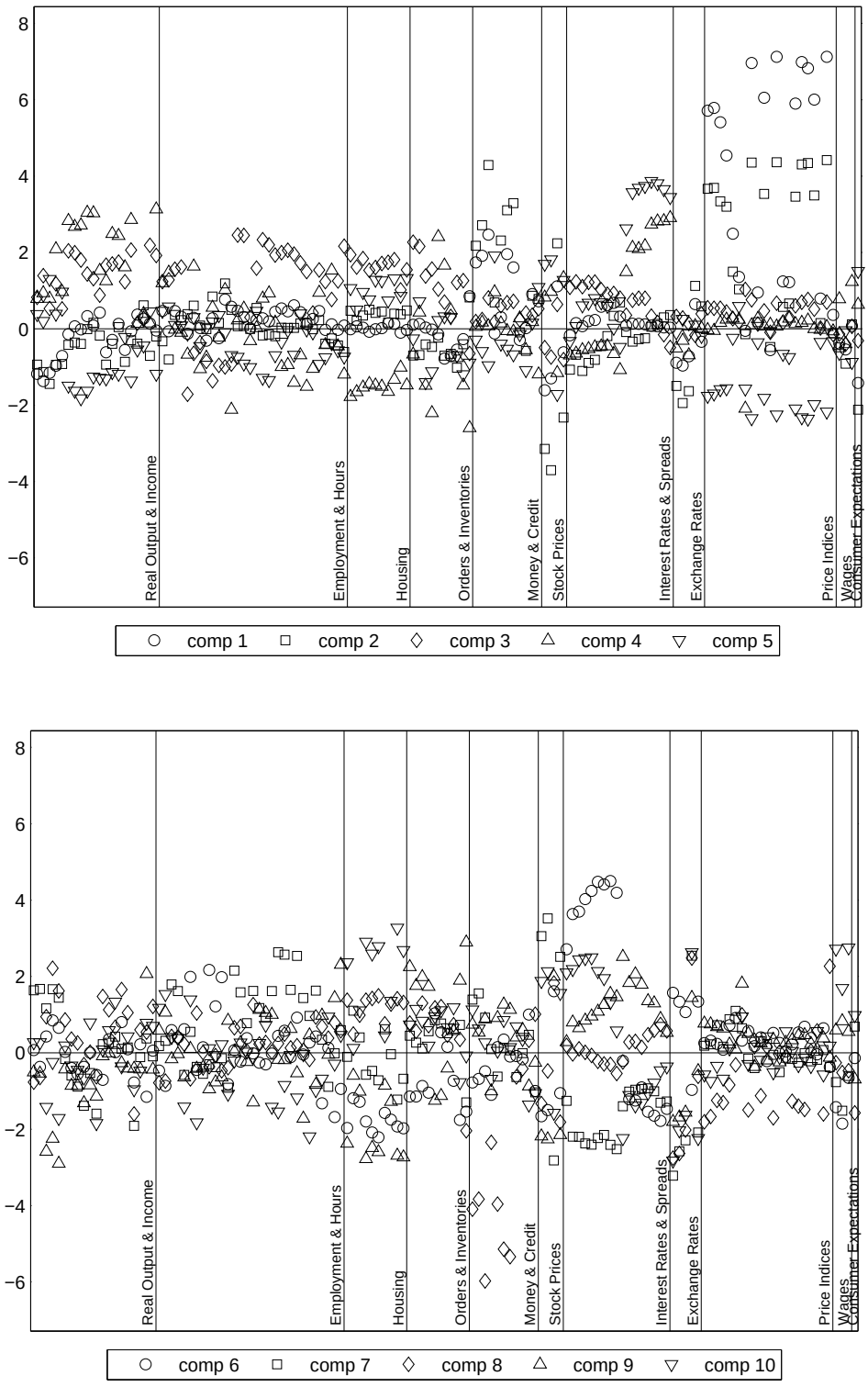

Figure 3: Nonzero factor loadings for the variables in the macroeconomic data set, using the L_2 criterion with $\lambda = 0$.

Figure 2 allows for a reasonable interpretation of the 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 2) 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.

Obviously, the interpretation of the factors remains subjective and often difficult. Nevertheless, sparsity helps. This is well illustrated for the variables in the group "Stock prices". Figure 2 shows that these variables only load on the 10th factor, all other factor loadings being zero. Using a non-sparse approach yields a much more diffuse pattern of loadings for this group, as we can see in Figure 3.

4.4 Forecasting Results

Using the 132 time series from the macroeconomic data set we forecast four key macroeconomic series; Industrial Production, Personal Income, Manufacturing & Trade Sales, and Employment. To quantify the forecast performance, we use a rolling window 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. For each series to forecast, the RMSE, the mean and median absolute forecast error are computed. The results are reported in Table 4.

First, we compare the forecast performance of the sparse methods ($\lambda > 0$) to their non-sparse counterparts ($\lambda = 0$). We see that there is hardly any difference in forecast performance. Adding sparsity does not yield a loss, but neither a gain in forecasting performance in this example. The simulation study already showed that the gain in forecasting precision, if the forecasting model is well specified, is modest. A first conclusion is that, while sparse factors are easier to interpret, they do not lose forecast performance.

Table 4: Forecasting Industrial Production, Personal Income, Manufacturing & Trade Sales, and Employment from the macroeconomic data set.

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
	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

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.

Secondly, we want to compare the relative performance of the three different criteria, L_2, L_1 and Tukey. For Industrial Production and Personal Income (Table 4), we find that robust methods often perform better than the benchmark of standard PCA, irrespective of which measure we use to evaluate the performance. The results for the other two series, Manufacturing & Trade Sales and Employment show that standard PCA forecasts perform well for these series. We can conclude that the presence of the outliers in this macroeconomic data set does not affect the performance of the standard PCA forecasts too much. Even if the estimated factors may be strongly influenced by the outliers, they still provide a diffusion index performing well for forecasting. However, as documented in the simulation study, there may be types of outliers where the L_2 approach is more vulnerable to outliers. While the robust estimators provide a safeguard with respect to outliers, they perform, on the whole, at least as well as the forecasting procedure based on standard PCA.

5 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 to investigate their potential for macroeconomic forecasts. Compared to standard principal component analysis, our proposed method gives a much closer approximation to the true factor space for heavy tailed error distributions or if outliers are present in the data. Imposing sparsity further reduces estimation error if the true factor structure is sparse, but, more importantly, provides easier to interpret loading matrices.

We considered two robust estimation criteria: a least absolute deviation loss function and the bounded Tukey biweight loss function. While the Tukey method provides even more protection with respect to outliers, in particular bad leverage rows, the L_1 approach preformed well in the empirical application. For the Tukey method, the loadings and factor scores are computed using a simple alternating iteratively reweighted least squares scheme. Alternating regression schemes have the advantage that they can cope easily with missing values in the data matrix.

If prior knowledge on a sparse factor structure is available, it is of course possible to impose a priori that certain elements of the loading matrix are zero. Also, if a natural grouping is present in the data, as is the case in the macroeconomic data set analyzed in Section 4 of this paper, the block structure of the variables can be taken into account in the factor construction procedure, as in Hallin and Liška (2011). The sparsity approach put forward in this paper does not require prior knowledge, but sets factor loadings to zero in a data-driven way. It can be used as an informal test to check whether prior assumptions are reasonable.

To conclude, we find that robust and sparse estimation of factor models has a great potential for improving both the interpretability of the estimated factors and the statistical accuracy in presence of model deviation. Developing robust estimators for related models, such as the dynamic factor model of Forni et al. (2005) or the Bayesian VAR model of Bańbura et al. (2010), is an open area for future research.

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 and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika*, 87:603–618, 2000.

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.

C. Dehon, M. Gassner, and V. Verardi. A Hausman-type test to detect the presence of influential outliers in regression analysis. *Economics Letters*, 105:64–67, 2009.

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.

M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100:830–840, 2005.

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

M. Hallin and R. Liška. Dynamic factors in the presence of blocks. *Journal of Econometrics*, 163:29–41, 2011.

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.

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.

V. Verardi and C. Croux. Robust regression in Stata. *Stata Journal*, 9:439–453, 2009.

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.