Abstract
This paper will introduce a new dimensionality reduction technique based on the ideas behind principal components analysis and the least absolute deviations estimator. We show that this technique can be used to obtain robust estimates of diffusion indexes for use in macroeconomic forecasting, thus eliminating the need to screen the data for outliers beforehand and thereby possibly providing a more efficient estimator. The performance of the estimator is tested using two Monte Carlo simulation studies, and a forecasting application of US macroeconomic series is provided. Overall the results are very encouraging.

Note: This is an unfinished draft and as such some of the sections are incomplete. In general we provide a full description of our estimation framework, whereas we only provide an outline of our preliminary results in the simulation and application sections.

Keywords: Forecasting, Factors Models, Principal Components Analysis

1. Introduction
As time goes by we accumulate information at an increasing rate, this coupled with the vast improvements in computation power this has driven a intensive research interest in the area of econometric analysis of large dimensional data sets. In macroeconomic forecasting in particular we often have many hundreds of time series at our disposal for forecasting, these are, however, of no interest without the right set of tools for analysing them.

Principal components analysis (PCA) is one of the key methods used for handling large data sets by means dimension reduction. PCA has two very appealing characteristics: it involves only fairly simple computations even when facing large data sets, and it is a well developed technique dating back to Pearson (1901) and Hotelling (1933) among others, and is thus now a common part of any textbook on multivariate statistics.

Over recent years PCA has also become quite used in the economic literature. One main reason for this is the fact that PCA can, under the right set of assumptions, provide consistent estimates of the factors in a factor model. The econometrics of principal component (PC) factor estimates have been developed through a number of significant contributions to the literature. One of the first investigations into the estimator was given by Connor and Korajczyk (1986, 1988) who considered consistency of PC factor estimates in a large $N$, fixed $T$ setup. Their results have since been extended into a large $N$, large $T$ setup, and for various assumptions on the error terms. The key references include Forni and Reichlin (1996, 1998), Forni et al. (2000), Stock and Watson (2002a), Bai and Ng (2002), Bai (2003)

This paper will focus on forecasting using PC factor estimates. We will use the so-called diffusion index (DI) forecasting methodology of Stock and Watson (2002a). The term diffusion index was chosen by Stock and Watson because they interpret estimated factors in terms of the diffusion indexes developed by NBER business cycle analysts. Building on their framework we will propose an alternative specification of the PC estimator which is more robust to outliers, and therefore aims at providing better macroeconomic forecasts.

The paper is outlined as follows. In Section 2 we give a quick overview of the DI forecasting methodology and the data set we use. Section 3 gives a short introduction to PCA and develops our least absolute deviations (LAD) inspired principal components estimator. Section 4 will evaluate the performance of this estimator compared to PCA using two simulation studies. Section 5 will test our estimators merits in a small empirical illustration. Finally, Section 6 gives the concluding remarks.
2. Diffusion Index Forecasting

Assume we have data on \( N \) possible predictors for \( T \) periods in a \( T \times N \) matrix \( X = (x_1, \ldots, x_T)' \), and that we are interested in forecasting \( y_{T+h}^{h} \), the annualised value of \( y_t \) in period \( T + h \). In the diffusion (DI) index forecasting framework of Stock and Watson (2002a) such a \( h \) period-ahead forecast would then be obtained by the following three steps.

First a factor model is estimated for the possible predictors:

\[
x_t = \Lambda f_t + \epsilon_t
\]  

this gives us a set of factor estimates which can be used as regressors in a forecasting equation. Hence, in the second step we estimate, by OLS, the forecasting equation using data for \( t = 1, \ldots, T - h \):

\[
y_{T+h}^h = \alpha_h + \beta_h(L)' f_t + \gamma_h(L) y_t + \epsilon_{T+h}^h
\]  

where \( f_t \) are the estimated factors. Having estimated this, we then simply, as the final step, fit the forecasting equation using the entire sample to obtain the desired forecast:

\[
y_{T+h|T}^h = \hat{\alpha}_h + \sum_{j=1}^{m} \hat{\beta}_{hj} f_{T-j+1} + \sum_{j=1}^{p} \hat{\gamma}_{hj} y_{T-j+1}
\]  

In order to be able to estimate the forecasting equation we, of course, first need to choose the number of factors; number of lagged factors; and number of lagged \( y_t \) to be included.

Note that the precise definition of \( y_t \) depends on the integration order of the variable of interest. We will use the same definitions as Stock and Watson (2002b). Hence if the variable of interest \( z_t \) is assumed I(1) in logarithms, then we have that

\[
y_t = 1200 \log(z_t/z_{t-1})
\]  

and if \( z_t \) is assumed I(2) in logarithms then

\[
y_t = 1200 \log(z_t/z_{t-1}) - 1200 \log(z_{t-2}/z_{t-1})
\]  

The main focus of this paper will be the estimation of \( f \). In the DI framework, as Stock and Watson described it, this would be done by means of Principal Components Analysis (PCA). However, PCA is unfortunately very sensitive to outliers. This was also recognised by Stock and Watson who applied a screening rule to the data in order to eliminate outliers beforehand. By their rule an observation exceeding the median of the series by more than 10 times the inter-quartile rule was defined as an outlier. Hence such observations would be deleted, and in cases where a balanced panel was needed/desired the entire series was dropped.

In this paper we propose a robust alternative to PCA for estimating \( f \), we show by simulation studies that in the presence of outliers it is superior to PCA and when outliers are not present the performance is very similar to that of PCA. This robustness does, however, naturally come at a cost, as our estimator is much more computationally expensive than PCA.

Before continuing to the description of the estimation methods, we will quickly comment on the data used.

2.1. Data

The data set used throughout this paper has been used a number of times in the literature. It is originally from Stock and Watson (2005), and has also been used by e.g. Bai and Ng (2008) and De Mol et al. (2006). The version used in this paper is updated one from Ludvigson and Ng (2009) which covers 131 macroeconomic series in the period 1964:1 to 2007:12.

We will consider forecasts of three different variables: Consumer Price Index (CPI), Industrial Production (IP), and Personal Income (PI). CPI is assumed to be I(2) in logarithms, and IP and PI are assumed to be I(1) in logarithms.

3. Principal Components and Least Absolute Deviations

PCA is a dimension reduction technique where we seek to express a large number of variables in terms of a smaller number of artificial variables. Suppose we have a dataset consisting of \( N \) variables with \( T \) observations for each variable, and collect these in a \( T \times N \) matrix \( X \), then the task is to find linear combinations of these \( N \) variables that best describe the underlying data, i.e. we are looking for components of the form \( f = X \lambda \), where \( \lambda \) is a \( N \)-vector of loadings associated with the component. What really defines PCA is the way we define “best”, here this will be measured in terms of variance, so we are looking for the component \( f \) which has maximal variance. Obviously such a factor will not be unique, and we will need to impose some identifying restrictions. We will therefore require that all components are uncorrelated and that the loadings \( \lambda \) have length 1.

More formally, PCA is concerned with finding a small number of components which solve the maximisation problem:

\[
\lambda_r = \arg \max_{\lambda} \text{var}(f) \quad \text{s.t. } \lambda_r' \lambda_r = 1, \lambda_r' \lambda_q = 0 \forall r \neq q
\]  

This will provide the loadings for the \( r \)th component which will have maximal variance conditional to being uncorrelated with all other components. In general we will let \( R \) denote the number of components of interest, and index these by \( r \). A very common assumption in the literature, which we also adopt, is that all the variables have been
centered. Under this assumption it is well known that \( \lambda_r \) is simply the \( r \)th eigenvector of \( X'X \), which has the corresponding eigenvalue \( \text{var}(f_r) \). So what appears a difficult problem is easily solved by determining the eigenvectors of \( X'X \). If \( X \) has rank \( R \), then \( X'X \) will have \( Q \) eigenvectors with non-zero eigenvalues, and thus \( X \) will have \( Q \) principal components.

An equivalent way of stating this problem is that we want to decompose \( X \) into \( R \) rank 1 matrices:

\[
X = \sum_{r=1}^{R} f_r \lambda'_r + E
\]

where \( E \) is a residual matrix. If we were to include all \( Q \) components in the sum there would be no residual, \( E = 0 \), and in this case we would have:

\[
X'X = \left( \sum_{r=1}^{R} f_r \lambda'_r \right)' \left( \sum_{r=1}^{R} f_r \lambda'_r \right) = \sum_{r=1}^{R} f'_r f_r \lambda_r \lambda'_r
\]

Comparing this with the spectral decomposition of \( X'X \) we see that \( \lambda_r \), represents an eigenvector of \( X'X \) and that \( f'_r f_r = \text{var}(f_r) \) is the corresponding eigenvalue. Thus clearly showing the equivalence with (8).

Of course estimating all \( Q \) components is not of interest since we are interested in dimension reduction, thus we will return to (9). Suppose that we were to disregard our knowledge of PCA and simply wanted to estimate the components in (9), a natural approach could then be to fit the parameters using least squares. In this case we would have a problem of the form:

\[
(\hat{F}, \hat{\Lambda}) = \arg \min_{F, \Lambda} (NT)^{-1} \sum_{n=1}^{N} \sum_{t=1}^{T} |x_{tn} - \sum_{r=1}^{R} \lambda_r f_r|^2
\]

\[
\text{s.t. } \Lambda' \Lambda = I_R
\]

where \( F = (f_1, \ldots, f_R) \), and \( \Lambda = (\lambda_1, \ldots, \lambda_R) \). This problem can be shown to be equivalent to (8).

Unfortunately, it is not easily solved. However, if we are only interested in the first component then this can be found quite easily by solving the problem for a single component only:

\[
(\hat{f}, \hat{\lambda}) = \arg \min_{f, \lambda} (NT)^{-1} \sum_{n=1}^{N} \sum_{t=1}^{T} (x_{tn} - \lambda f_t)^2
\]

\[
\text{s.t. } \lambda^2 = 1
\]

Once this has been found it follows from (9) that the residual matrix from the regression, \( E = X - f^\lambda \), will have rank \( R - 1 \), and that \( E' E \) will have a largest eigenvalue corresponding to the second largest eigenvalue of \( X'X \). Thus if we simply solve (12) using the residual matrix \( E \) instead of \( X \) we will obtain the second component. This iterative scheme can be applied repeatedly to find all components. Hence we have the following algorithm to find the principal components iteratively by direct application of least squares:

**Algorithm 1. Iterative approach to solving (11).**

(i) Solve the minimisation in (12) disregarding the condition that \( \lambda^2 = 1 \) to find the unrestricted parameters \( \lambda^u \) and \( f^u \).

(ii) Scale the parameters such that the condition \( \lambda^2 = 1 \) is satisfied, i.e., let \( \lambda = \lambda^u / \sqrt{\lambda^u \lambda^u} \), and \( f = f^u \sqrt{\lambda^u / \Lambda^u} \).

(iii) Compute the residuals of the regression: \( E = X - f^\lambda \).

(iv) Start over using \( E \) instead of \( X \) in (12) to find the next component.

Although we have, using PCA, a simple way of finding the components of (9), it could very well be interesting to consider alternatives to least squares for estimating these. The focus of this paper is robustness to outliers and as such there is a very natural alternative to least squares, namely least absolute deviations. Reformulating (11) as a LAD problem we get:

\[
(\hat{F}, \hat{\Lambda}) = \arg \min_{F, \Lambda} (NT)^{-1} \sum_{n=1}^{N} \sum_{t=1}^{T} |x_{tn} - \sum_{r=1}^{R} \lambda_r f_r|^2
\]

\[
\text{s.t. } \Lambda' \Lambda = I_R
\]

Unfortunately, this is an even harder problem to solve, and therefore we will apply the same iterative scheme as in the least squares, and further instead of solving (13) for a single component directly, we will approximate the objective function. The true objective function for a single component is:

\[
V_{\text{LAD}}(f, \lambda) = (NT)^{-1} \sum_{n=1}^{N} \sum_{t=1}^{T} |x_{tn} - \lambda f_t|^2
\]

Normally, such a problem would be solved by means of linear programming. This problem, however, is not linear in the parameters and thus we need to consider alternative solution methods. We propose to smooth the objective function in a similar fashion as the non-linear smoothed LAD estimator by Hitomi and Kagihara (2001):

\[
V_{\text{LAD}}^d(f, \lambda) = (NT)^{-1} \sum_{n=1}^{N} \sum_{t=1}^{T} \sqrt{(x_{tn} - \lambda f_t)^2 + d^2}
\]

where \( d > 0 \). Recall that \( |z| \approx \sqrt{z^2 + d^2} \). This is, however, not differentiable. Thus we use the approximation \( |z| \approx \sqrt{z^2 + d^2} \). Since \( \sqrt{z^2 + d^2} \leq \sqrt{z^2 + d} = |z| + d \), the approximation will at most be a distance of \( d \) from the true function. We therefore have for our problem that:

\[
0 \leq V_{\text{LAD}}^d(f, \lambda) - V_{\text{LAD}}(f, \lambda) \leq d \quad \forall \lambda, f \]

Hence, if we choose \( d \) appropriately, then the difference between the two objective functions will disappear asymptotically. An obvious choice for \( d \) could be \( d = (NT)^{-1} \).
In the remainder of this paper we will let LAD factors denote factor estimates obtained using the approximate objective function (15) solved iteratively using Algorithm 1 with \( d = (N T)^{-1} \). Our aim is thus to investigate how these compare to the traditional PC factor estimates.

4. Monte Carlo Simulation

In order to compare the performance of the two methods we will start by conducting two simulation studies. The first will investigate the effect outliers have on factor estimation, and the second will assess how these effects propagate to the DI forecasts.

The way we will generate outliers is strongly motivated by the outlier detection rule used in Stock and Watson (2002b) as described in Section 2. Let \( x_{tn} \) be an observation in the original (possibly simulated) data set, then the corresponding “noisy” data set where we have added outlier in the original (possibly simulated) data set, then

\[
\tilde{x}_{tn} = \begin{cases} 
  x_{tn} & \text{w. prob. } 1 - p \\
  x_{tn} + a \text{IQR}(x_{n}) \text{sgn}(x_{tn}) & \text{w. prob. } p
\end{cases} \tag{17}
\]

where IQR(\( x \)) is the inter-quartile range of \( x \), and \( \text{sgn}(\cdot) \) is the sign function. Using this specification, the noise is controlled by two parameters, \( a \) which determines the size of the outlier and \( p \) which is the probability with which it will occur. Our outlier generation rule is clearly very closely related to Stock and Watson’s outlier detection rule, and will ensure that if we choose \( a < 10 \) the fraction of outliers in the “noisy” data set will be approximately zero when measured by their rule.

We consider four different cases of outliers, where we vary the parameters of the outlier generation rule. These are: Case 1: \( a = 15 \), \( p = 0.05 \); Case 2: \( a = 5 \), \( p = 0.05 \); Case 3: \( a = 15 \), \( p = 0.01 \); Case 4: \( a = 5 \), \( p = 0.01 \). Thus we have two cases, 2 and 3, where Stock and Watson’s outlier detection rule would not detect (and remove) the generated outliers.

4.1. Factor Estimation

In this first simulation study we want to consider how precise the two methods estimate the distribution of the factors in a factor model when outliers are present. We will assume that the data are generated from the factor model (1), with three factors, \( f_r \), i.e. \( R = 3 \). Their distributions will be

\[
\begin{align*}
  f_1 & \sim N(0, 30) \tag{18} \\
  f_2 & \sim N(0, 20) \tag{19} \\
  f_3 & \sim N(0, 10). \tag{20}
\end{align*}
\]

The factors have been chosen such that the variance of the first factor is larger than that of the second factor which in turn is larger than that of the third factor. This ensures that the methods will estimate the factors in this particular order. The factor loadings will be given as:

\[
\Lambda = \sqrt{R/N} I_R \otimes \mathbf{t}_{N/R}, \tag{21}
\]

where \( \mathbf{t}_k \) is a \( k \)-vector of ones. In other words each factor will get an equal “share” of the variables and the loadings for these variables will all be equal. Furthermore this specification is in accordance with the identifying restriction \( \Lambda' \Lambda = I_R \), which obviously is necessary in order to be able to pick up the true distributions. The error component of the variables will be chosen to be standard normal, i.e. \( \epsilon_i \sim N(0, 1) \).

In Figure 1 we have plotted the distribution of the first factor when we pool the estimates across simulation runs in the four cases of outliers. The graphs include both the PC and LAD estimates as well as the true distribution/data-generating process (DGP). A note of caution should be given here: the graphs have been created using a kernel-based smoother, and should therefore be interpreted with some caution.

We clearly see that in all cases the LAD estimates are much closer to the DGP than the PC estimates. The effect of the outliers is clearly more pronounced as \( a \) and/or \( p \) increase. However, also in the case when \( a = 5 \), i.e. when Stock and Watson’s rule would not detect the outlier, we see that there is an effect.

To get a more complete picture of the effects, Table 1 provides the first four sample moments of the estimated factors. Here we see the same effects as in the figure. The variance is very affected by the outliers in the PC estimates, this is not the case for the LAD estimates. Moreover, in some cases, the kurtosis is also quite affected in the PC estimates.

These findings suggest that when the data set contains outliers of a sufficient size or in sufficient numbers then PC will not correctly estimate the factors. Although the LAD estimates are also affected this is to a much lesser degree, and as such they appear to provide robust estimates of the factors.

4.2. Forecasting

The second simulation study will assess the performance of DI forecasts obtained using the two methods for estimating the factors. We will do this using real data instead of postulating a data-generating process and using this to generate data. A subset of the data described in the previous section is used and then outliers are generated in the same fashion as the first simulation study according to the four cases.

For the variables of interest we estimate the model for the period 1990:1 to 2006:12. This model is used to obtain in-sample 12-month ahead forecasts for the period 1992:1 to 2007:12. The DI forecasting equation will be estimated with four factors, four lags of these, and six lags of \( y_t \).

The simulation is done in the following way. For each of the cases of outliers we generate a “noisy” data set 100
Figure 1: Plots of the estimates of factor 1 using the two methods along with the data-generating process for cases 1–4 (left to right). Note that the estimates have been pooled across simulation runs in order to obtain a single density estimate.

Table 1: Averages of the first four sample moments of the estimated factors using the two methods. The table provides results for the four cases of outliers as well as for a reference case with no outliers.

Table 2: Average mean squared forecast errors for the four cases of outliers and as reference the mean squared error of forecasts obtained using the true data.

In Figure 2 the four sequences of MSEs for the different cases of outliers are plotted both in the case of the PC and the LAD estimates. Here we see in more detail why the average MSE is higher for the PC estimates of CPI. When using LAD estimates the MSEs are almost identical regardless of outliers. However, when we use the PC estimates, especially in the end of the period, we see that the MSEs deviate quite a lot from the reference case. These findings suggest that robust estimation of the factors carries over to the forecasting equation in providing robust forecasts.

5. Empirical Illustration

We end this paper by giving a small empirical illustration. We will use the data set described in Section 2 to conduct a simulated real-time forecasting experiment. The estimation period starts in 1964:1 and the first simulated out of sample forecast is made in 1974:1, i.e. we forecast 1974:1+h. We then expand the estimation sample by one period such that it ends in 1974:2, reestimate the model and forecast 1974:2+h. We continue by this approach until we reach the end of the data set. The results are summarised in Tables 3 and 4.

In general we see that regardless of the DI specification or the forecast horizon the relative MSEs are very similar.
Figure 2: In-sample forecasts of CPI for the period 1992:1 to 2007:12. The top panel is estimated using PC factors, and the bottom panel is estimated using LAD factors. In the four cases of outliers, the lines represent mean squared forecast errors, where the mean is taken across 100 simulation runs. The reference case shows the realised squared forecast errors from the true data set. The DI equation includes four factors, four lags of these, and six lags of \( y_t \).

6. Concluding Remarks

In this paper we have proposed a robust alternative to PCA and demonstrated its use in the diffusion index forecasting framework. We have shown that in the presence of outliers it is superior to PCA and can be used to obtain robust forecasts.

The work in this paper is still at an early stage, however, based on the preliminary results we believe that our method can become a useful tool in macroeconomic forecasting. There are, however, still a number of open issues. Among these are the investigation of the statistical properties of the estimator and the extension of the method to include estimation using unbalanced data sets.
References


Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. Econometrica 70 (1), 191–221.


Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6 2 (11), 559–572.

