State University – The Higher School of Economics

Economics Faculty

Master Program «Mathematical Methods of Economic Analysis»

Mathematical Economics and Econometrics Department

MASTER THESIS

«Using Principal Covariate Regression for Macroeconomic Time Series Forecasting: Comparative Analyses based on the Monthly U.S. Data»

Group № 71MMEA
Bulavskaya T.Y.

Supervisors:
Demidova O.A.
Associate professor

Dick van Dijk
Professor, Econometric Institute
Erasmus University Rotterdam

Moscow 2010

Abstract

In the presented master thesis a problem of forecasting U.S. real economic variables growth rates by the means of dynamic factor models is considered. Forecasting horizons vary from 1 month to 1 year. The research is focused on different methods of dynamic factors' estimation. The following modifications of the standard approach are investigated. Firstly, implementation of analysis and selection of predictor variables prior to a factors' estimation step. Secondly, use of principal covariate regression instead of more standard principal component regression. Thirdly, consecutive use of variables selection and principal covariate regression methods. Forecasting accuracy conclusions are based on comparison of mean squared prediction errors and recession periods dating. Empirical results stand for introduced modifications and their combination.

Table of contents

Introduction

This work studies the question of forecasting business and economic activity in case of accessibility of a wide range of economic variables, which could be used as predictors. Traditionally academic papers concerning macroeconomic modeling and forecasting are based on quite parsimonious models with the limited number of explanatory variables suggested by economic theory. However many businessmen and political authorities have already realized the benefits of using and tracking down dynamics of a large number of variables for real-time decision making.

By the end of the past century this subject has become of increased interest, particularly because of the accessibility of high dimensional datasets. For example, in the U.S. information on more than a thousand economic indicators is available, and each of them could be considered as a potential signal of economic development. But use of a large number of explanatory variables changes a forecasting procedure. If the number of predictors is commensurable with the number of observations, ordinary least squares results in overfitting and, consequently, in a highly dispersed forecast. In addition, it is more likely to detect multicollinearity in a high dimensional dataset. Forecasts of higher quality could be obtained by using only key information, preliminary extracted from all the variables available.

In order to emphasize the cyclical component of the variables, but at the same time to diminish individual noisy components, separate variables are combined into composite indexes. Indexes are usually constructed as a linear combination of observed variables.

The Conference Board applies a so-called “non-model-based” approach to construct composite coincidence and leading indexes in real time. This approach consists of employing certain weighting rules. “Model-based” approaches include Markov-switching models and Dynamic factor models. The

Chicago Fed National Activity Index[1] (USA) and European Coincident Index[2] (Europe) are well-known examples of published factor-based indexes.

On one hand, the "non-model-based" approach is easily implemented and interpreted and doesn't suffer from overfitting problem. On the other hand, there are several disadvantages: weighting scheme is time-invariable, lag values are not taken into account and explicit link to a target variable is absent.

Here we consider a modification of dynamic factor models which allows to minimize the disadvantages of the "non-model-based" approach. Factor models are based on the hypothesis that a big group of observed economic variables is affected by a limited number of common trends and individual idiosyncratic shocks. The most popular way of extracting these common trends (factors) is principal components analysis. But a practitioner faces a number of questions. Do all the available economic variables are relevant for forecasting a certain target variable/index? How does one connect a factor extraction step with the final aim of modeling – forecasting?

To overcome these problems one can preliminary select relevant variables and/or use principal covariate analysis. In this work we offer a description of these two approaches and a comparison of empirical forecasting results using the methods separately and sequentially. Hence the main research question of this work is:

Can we increase forecasting accuracy of macroeconomic time series by sequential application of predictor variables selection and principal covariate analysis?

The work is organized as follows. A brief literature review is presented in Chapter 1. In Chapter 2, relevant methodology is described. Chapter 3 consists of a data description, forecasting procedures and instruments used to

---

[1] For more details see: http://www.chicagofed.org/webpages/research/data/cfnai/current_data.cfm
[2] For more details see: http://eurocoin.cepr.org/.

evaluate the relative performance competing forecasting models. In Chapter 4, results of empirical forecasts are presented. Chapter 5 concludes on the main empirical results.

Chapter 1. Literature review

As soon as one have decided on target variables for forecasting, explanatory, or leading, variables have to be determined. In this Chapter we review recent academic literature concerning explanatory variables selection. We go into the details of choosing target variables in the next Chapter.

1.1. Dynamic factor models

The methods of dynamic factor models and principal component regression were thoroughly studied in both theoretical and empirical researches. It has acquired a reputation of being the most successful in forecasting macro data.

In the paper «Macroeconomic forecasting with diffusion indexes» Stock and Watson (2002) aim to forecast eight monthly U.S. macroeconomic time series using 215 leading independent variables. Authors investigated the forecasting ability of diffusion indexes over 6, 12 and 24 months ahead. Total number of observations is about 480. The target variables are: the industrial production index, personal income less transfers, manufacturing and trade sales, total nonagricultural employment, the consumer price index, the personal consumption expenditure implicit price index, the consumer price index less food and energy, the producer price index for finished goods. As follows from the title authors evaluated the ability of dynamic factors, or how they called them “diffusion indexes”, in real and price economic indicators forecasting. As benchmark models they used univariate autoregression, trivariate vector autoregression and autoregression with distributed lags. Different forecasts were compared basing on the mean squared prediction error (MSPE). Stock and Watson showed that the dynamic factor models substantially excel traditional econometric models in terms of forecasting accuracy. Real economic variables require up to five factors, price indexes are well predicted by only one factor and lags of the target variable.

Analysis of the extracted factors as well deserves attention. Firstly, authors found out that the first six principal components explain in average 39% of original variables' variation. The first twelve components explain more than a half of original variation. As Stock and Watson accentuate, this result illustrates the hypothesis of a limited number of reasons explaining macroeconomic fluctuations. Secondly, authors analyze correlations between the factors and the observed variables and conclude:

- The first factor mostly transmits dynamics of output and employment variables;
- The second factor – spreads, unemployment and inventories variables;
- The third factor – interest rates variables;
- The fourth factor – stock prices variables;
- The fifth factor – price indexes variables;
- The sixth factor – housing starts and sales variables.

In 2005 Stock and Watson published an extended version of their research. The working paper is entitled «An empirical comparison of methods for forecasting using many predictors». The main differences are:

- A wider range of competing models were used. For example, authors included empirical Bayes and weighted and generalized least squares methods.
- The number of leading variables was reduced from 215 to 132.
- 60 additional observations were included.

The main conclusions of the working paper correspond to the ones in the article of 2002. Firstly, factor models allowed for more accurate forecasts, especially for short horizons. Secondly, all the three least square methods (OLS, WLS, and GLS) ended up with similar results and were not beaten by other models. Thus, Stock and Watson showed that factor models are

persistently good in relatively accurate forecasting, in spite of reduction of the independent variables and extension of the competing models.

Successful use of the dynamic factor models were established as well for more complicated models in evaluating monetary policy effects. Although this topic is not explicitly related to our research, it demonstrates **versatility** of the factor models. He we exemplify one paper published in 2005 by Bernanke, Boivin and Eliasz and entitled «Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach». According to Bernanke et al. the use of vector autoregressions for identification and measurement of the effects of monetary policy on macroeconomic indicators is entailed with the three following problems:

- Not all the information available to central banks and private sector in reflected in VAR models;
- Certain arbitrariness is presented in the process of choosing variables standing for the level of “real economic activity”.
- Impulse response functions are computable only the variables explicitly included in the model.

Authors suggested using the principles on which dynamic factor models are based to confront these problems. They used two estimation methods for FAVAR:

- Two-step-approach: 1) extract factors by means of principal components; 2) use estimated factors as independent variables in VAR.
- One-step-approach: simultaneously estimate factors and VAR parameters using Bayesian likelihood methods and Gibbs sampling.

Authors conclude that FAVAR indeed extracts the most relevant information from a big number of predictor variables. Also Bernanke et al. obtained an impulse response function for each of the exogenous variable. Authors used them to evaluate empirical reasonability of FAVAR models specifications. Most of the impulse response functions were of expected sign and magnitude. Authors as well showed that increasing the number of estimated factors from one to five doesn't give qualitative change in the results. Point estimates of the two mentioned estimation approached are in general coincide. But the Bayesian method gives a much higher estimates' variation. Authors attribute it for an excessively strict structure of likelihood estimation.

Among the most significant works in the area of DFM are Croux, Renault, and Werker (2004), Forni, Hallin, Lippi, and Reichlin (2000, 2003, 2004), Bai, and Ng (2002, 2006), Boivin, and Ng (2006), Moench (2008). Stock and Watson (2006), Marcellino (2006) review a wide range of the latest papers concerning forecasting with many predictors.

1.2. Variables selection

In the considered works authors do not give a due consideration to selection of the leading variables. Of late years there are more empirical papers questioning the use of all the available variables for accurate forecasting. Frequently researchers tend to use the same set of leading variables to forecast completely different target variables. Principal components analysis aims to maximize just the variance of extracted factors[3]. So if some of the exogenous variables have no predicting power for the target variable, factors are "noisy". Situation is aggravated if "noisy" variables are presented in a highly correlated group.

In the paper «Forecasting economic time series using targeted predictors» Bai and Ng (2008) introduced a notion of targeted predictors and suggested bringing two improvements in dynamic factor models:

- Nonlinear principal components analysis.
- Preliminary reduction of the number of predictor variables used for factor estimation through previously determined selection procedures.

[3] See Chapter 2 Methodology for the method refreshment.

Here we discuss only the second improvement. In Chapter 2 a more technical description of the procedures is presented. It should be noted that selection criteria are based solely of statistical properties of target and leading variables.

Authors predicted the consumer price index with the set of predictors from Stock and Watson (2005). Bai and Ng concluded that forecasting errors for different horizons (from 1 to 24 months) decrease substantially when only selected targeted predictors were used for factor estimation. In addition, authors showed that some groups of exogenous variables were selected systematically. But on the other hand, a set of targeted predictors changed with forecast horizon and certain sample. Therefore, the general practice of using a **fixed set** of predictor variables **constrains** a dynamic factor model.

For one of the latest studies, but more methodology specialized, we refer to Gelper and Croux (2008).

1.3. Principal covariate regression

Selection procedures are not the only way to find a link between dynamic factors and target variables. Heij, Groenen and Van Dijk (2007) in the paper entitled «Forecast comparison of principal component regression and principal covariate regression» considered a completely different approach. Authors presented the method proposed by De Jong and Kiers (1992) – Principal Covariate Regression – but adopted for a time series application. This method avoids the two-step-procedure of "classical" dynamic factors models. But with it authors noted: «…PCovR is a data-based method that does not employ an explicit underlying statistical model. As the construction of the PCovR factors is directly related to their use in forecasting, this may give better forecasts as compared to two-step methods like PCR.»[4]

Authors examined proposed method on both simulated and empirical data. In simulation example different data generating processed were

[4]Heij, Groenen and Van Dijk (2007), p. 3613.

considered. Empirical part included 12-months-ahead forecasting of 4 real variables from Stock and Watson (2002). Firstly, principal covariate regression gave more accurate forecasts with the **less** number of factors. Secondly, mean squared prediction error was reduced by maximum 50%. The most substantial decrease was observed for the industrial production index and manufacturing and trading sales.

Summarizing the chapter, the main conclusions of the reviewed papers indicate superiority of the dynamic factor models in different modifications over the standard econometric forecasting models. It was shown, that the models work for forecasting different target variables, on different samples and with different sets of leading variables. Meanwhile the number of academic papers using standard principal components for factors estimation is quite big. But it was not yet given a deserved consideration to independent variables selection and principal covariate regression.

Further presented research is based to a large extent on the works of Heij, Groenen and Van Dijk (2006, 2007) and was carried out with the support and advising of the authors.

Chapter 2. Methodology

In this chapter a technical description of dynamic factors' estimation methods and forecasting models is presented. This work considers only single-factor models. A single factor is interpreted as an integral leading index of a target variable.

Firstly, turn to the notation conventions:

Y - a vector of size *T*, consisting of the target variable values over a given period of time. We assume that the series is already stationary. If any additional transformations are need, we will specify it.

y_t - value of the target variable at the moment *t*.

h - forecasting horizon.

$\hat{y}_{t+h}$ - forecasted value of the target variable at the moment *t*+*h* based on the information available at the moment *t*.

Z - a matrix of size *Txk*, consisting of the preferential predictors values over a given period of time. Preferential predictors are always included into a forecasting model due to their economic interpretation. We assume that all of the series are already stationary. Here *k* is a number of different preferential predictors.

z_{ti} - value of *i*-th preferential predictor at the moment *t*.

X - a matrix of size *Txn*, consisting of the leading predictor variables values over a given period of time. Preferential predictors and constant are not included into matrix X. Each of the predictors is assumed to be stationary. Here *n* is a number of different leading predictors.

x_{ti} - value of *i*-th leading predictor at the moment *t*.

F - a vector of size *T*, consisting of the single dynamic factor values over a given period of time.

f_t - value of factor at the moment *t*.

So far we are not specifying a method of the factor construction. We just mention that it is defined as a linear combination of the observed leading variables:

$$f_t = \sum_{j=1}^{n} \alpha_j x_{tj} , \tag{1}$$

for $t = 1,\ldots,T$.

Or in a matrix form:

$$F = XA , \tag{2}$$

where A is a vector of size *n*, consisting of elements α_j for $j = 1,\ldots,n$.

Relation between future values of the target variables and current and lagged values of the preferential predictors and factor is also assumed to be linear. Coefficients are estimated from the following time series model:

$$y_{t+h} = \alpha + \sum_{j=0}^{r} z_{t-j\cdot}\beta_j + \sum_{j=0}^{q} f_{t-j}\gamma_j + \varepsilon_{t+h} , \tag{3}$$

where *r* and *q* are numbers of included lags of the preferential predictors and factor; $z_{t\cdot}$ is a row-vector of size *k*, consisting of all the preferential predictors values of the moment *t*, f_t is a value of the single factor at the moment *t*; α is a constant, β_j is a coefficient vector of size *k* for $j = 1,\ldots,r$, γ_j is a coefficient scalar for $j = 1,\ldots,q$. Numbers of included lagged values *r* and *q* are defined according to the smallest outcome of the Schwarz Information Criterion (SIC) for $r \le 5$ and $q \le 2$. An error term ε has a zero expectation value; errors are not mutually correlated and not correlated with the preferential predictors and factor. As soon as coefficients in the equation (3) are estimated, a forecasted value of the targeted variables is written as:

$$\hat{y}_{T+h} = \hat{\alpha} + \sum_{j=0}^{r} z_{T-j\cdot}\hat{\beta}_j + \sum_{j=0}^{q} f_{T-j}\hat{\gamma}_j . \tag{4}$$

The described forecasting approach is the most common one in the dynamic factors methodology. We refer to Stock and Watson (2002) for a more detailed manual.

In the following parts of the Chapter we consider several methods of estimating vector A in the equation (2) and coefficients of the model (3), and procedures of preselecting exogenous variables form the matrix X.

2.1. Factors as principal components

If the number of different exogenous variables (n) is proportionate to the number of observations (T), estimates and forecasts, obtained by OLS are unreliable. Therefore we are trying to transfer as much as possible information contained in X into one or several integral unobserved factors. In econometric models one explains **variability** of one variable through **variability** of other variables. So traditionally we consider **variance** and covariance of variables $x_{\cdot i}$ as a numerical measure of contained information. In other words, we want latent factors defined by the equation (1) to explain as much variance of $x_{\cdot i}$ as possible. It is easy to show that principal components of the data matrix X satisfy this requirement. But additional data standardization needs to be done before computing. Firstly, each column of the matrix X should be centered. It simplifies further computations, but do not influence forecasting results since all dependences are linear. Secondly, each column of the matrix X is normalized to have a unit variance[5]. This kind of standardization reduces the risk that the variable with a larger variance has a larger influence on principal components calculation than the one with a smaller variance. One can point out both positive and negative sides of this approach. The main advantage is that we smooth spread of observed variables caused by specific unit of measurement or by features of a certain variable. The main disadvantage is forced equalization of different variables for factors estimation and further forecasting[6].

One way to compute principal components is to calculate eigen values and eigen vectors of the variance-covariance matrix of the data X. Due to

[5] Here and further we use unbiased estimate of a sample variance.
[6] One can use weighted principal components to avoid this problem, see Boivin and Ng (2006).

standardization and normalization the matrix is written as $C = X' * X / (T-1)$, where X is transformed. Eigen vector corresponding to the largest eigen value is a vector A in the equation (2). But here we consider an alternative computation method – through a singular decomposition of the data matrix X. It is necessary for description of principal covariate regression. In the following description we consider a more general multi-factor model keeping logic and notation of Heij, Groenen and Van Dijk (2007).

Instead of maximizing variance of estimated p factors, one can consider a problem of approximation of the matrix X by the matrix $\hat{X}$ with rang p (in case of a single-factor model p equals one). In other words, the following Frobenius norm is minimized:

$$\left\| X - \hat{X} \right\|_F^2 = \sum_{t=1}^{T} \sum_{i=1}^{n} (x_{ti} - \hat{x}_{ti})^2 \Rightarrow \min_{\hat{X}}, \tag{5}$$

s.t. $rank(\hat{X}) = p$.

Matrix X can be presented as a singular value decomposition:

$$X = U * S * V', \tag{6}$$

where U is an orthogonal matrix of size *TxT*, consisting of the left singular vectors; S is a diagonal matrix of size *Txn*, consisting of the singular values ordered in a decreasing order; V - is an orthogonal matrix of size *nxn*, consisting of the right singular vectors.

According to Eckart-Young theorem the solution of the problem (5) is:

$$\hat{X} = U_p * S_p * V_p', \tag{7}$$

where U_p are the first p columns of the matrix U, S_p is a diagonal matrix of size *pxp*, consisting of p largest singular values of the matrix X ordered in a decreasing order, V_p are the first p columns of the matrix V.

Required approximated matrix could be written as:

$$\hat{X} = X * V_p * V_p'$$ [7]. (8)

To estimate latent factor we need firstly estimate matrix A in the equation (2). If we set matrix A as a matrix $V_p S_p^{-1}$ and matrix B as a $S_p V_p'$, than:

$$\hat{X} = X * A * B,$$

$$XA = U * S * V' * V_p * S_p^{-1} = U_p S_p S_p^{-1} = U_p,$$

$$F' * F = A' * X' * X * A = U_p' * U_p = I_p.$$

Since the right singular vectors of matrix X are identical to the eigen vectors of the matrix C accurate within constant multiplication, derived factors are the same as principal components.

Therefore, matrix A is a solution of the following extremal problem:

$$\|X - XAB\|_F^2 \Rightarrow \min_{A,B}, \quad (9)$$

s.t. $A'X'XA = I_p$.

As soon as matrix A is estimated, we are back to a forecasting problem of the target variable y according to the model (3). Parameters are estimated by the means of OLS:

$$\left\| y - \alpha - \sum_{j=0}^{r} Z(-j)\beta_j - \sum_{j=0}^{q} X(-j)A\gamma_j \right\|^2 \Rightarrow \min_{\alpha,\beta_j,\gamma_j}, \quad (10)$$

where $\| \ \|$ is a Euclidian vector norm, $Z(-j)$ is a matrix of lagged values of the preferential predictors, $X(-j)$ is a matrix of lagged values of the leading variables. When describing the problem (10) we slightly abused the notation. Since lagged values are used, first several observations of the target variable should be thrown away. We assume that sizes of the matrices in the problem (10) allow to compute the value of the objective function.

[7] Using singular value decomposition: $U_p * S_p * V_p' = U * S * V' * V_p * V_p'$. Since $V' * V_p$ is a block matrix of size *nxp*, consisting of an identity matrix *pxp* and a zero matrix *(n-p)xp*, $U_p * S_p = U * S$ by definition of matrices U_p и S_p.

2.2. Factors as principal covariates

When factors are computed as principal covariates two independent extremal problems (9) and (10) are combined. A weighted average of the two objective functions is minimized. For given weights ω_1 and ω_2:

$$\omega_1 \left\| y - \alpha - \sum_{j=0}^{r} Z(-j)\beta_j - \sum_{j=0}^{q} X(-j)A\gamma_j \right\|^2 + \omega_2 \| X - XAB \|_F^2 \Rightarrow \min_{A,B,\alpha,\beta_j,\gamma_j} \quad (11)$$

s.t. $A'X'XA = I_p$

Eventually we are interested only in relative weights of two summed objective functions. But since ω_1 is a weighting coefficient of the Euclidian vector norm and ω_2 - of the Frobenius matrix norm, we need to scale them:

$$\omega_1 = \frac{\omega}{\|y\|^2}, \quad (12)$$

$$\omega_2 = \frac{1-\omega}{\|X\|_F^2}, \quad (13)$$

where $0 \le \omega \le 1$, otherwise one of the weights is negative and the problem (11) has no solution. But the question of choosing ω is still open. One can use information criteria or cross-validation for this purpose. In the empirical application cross-validation was used with the following grid: 0.01, 0.1, 0.3, 0.5, 0.7 and 0.9.

The problem (11) could not be solved analytically due to nonlinearity: the same matrix A is multiplied by the leading predictors matrix X, all its lags $X(-j)$ $j = 1,...,q$ and matrices inside the Frobenius form.

De Jong and Kiers (1992) offered a static version of principal covariate regression without preferential predictors Z and lags $X(-j)$:

$$\omega_1 \| y - \alpha - XA\gamma \|^2 + \omega_2 \| X - XAB \|_F^2 \Rightarrow \min_{A,B,\alpha,\gamma}. \quad (14)$$

This problem is also non-linear but it could be solved analytically, see. De Jong, Kiers (1992). Heij, Groenen and Van Dijk (2007) demonstrated how a similar result could be achieved with two sequential singular value decompositions.

In the report «Time series forecasting by principal covariate regression» (2006) Heij, Groenen and Van Dijk suggested an algorithm for solving the dynamic problem (11). Hereafter we adduce the main steps and results of the algorithm.

Suppose a matrix A is known. Then the problem (11) turns into a linear one and could be estimated by the means of OLS. But still the value of the objective function depends on the matrix A. Values of the objective function could be found for all the possible matrices A of size *nxp* satisfying the condition $A'X'XA = I_p$. Therefore the problem (11) reduce to minimizing some nonlinear function $f(A)$ over a matrix A. The problem could be solved by the means of iterative majorization[8]. The algorithm of iterative majorization consists of the following steps:

1. Choose some initial matrix A_0, for example estimated by the means of principal components analysis.
2. Find a function $g_0(A)$ with the properties:
    a. $g_0(A) \geq f(A)$ for all A;
    b. $g_0(A_0) = f(A_0)$;
    c. $g_0(A)$ could be minimized analytically.
3. Solve a problem $g_0(A) \Rightarrow \min_A$, $A_1 = \arg\min_A g_0(A)$. Since the function $g_0(A)$ majors the function $f(A)$, then $f(A_1) \leq g_0(A_1) \leq g_0(A_0) = f(A_0)$. Thus if $g_0(A_1) < g_0(A_0)$, then $f(A_1) < f(A_0)$[9].
4. Go to step 2, increase all the subscripts by one.

Therefore we have a sequence of matrices $A_0, A_1, A_2,...$ and a corresponding decreasing sequence of the objective function values $f(A_0), f(A_1), f(\mathrm{A}_2),...$. However it should be noted that the algorithm does not

[8] See Kiers(1990) for more information on iterative majorization.
[9] In case of equality one can choose different initial matrix or different majorant function.

guarantee a global minimum, so iterations should be repeated for several initial points.

A majorant function could be chosen as $g_i(A) = \upsilon - 2tr(A'VA_i)$, where υ and V are computed using the target variable, preferential and leading predictors values and A_i. For more detailed explanation of the functions $g_i(A)$ we refer to the report Heij, Groenen and Van Dijk (2006).

2.3. Targeted predictors

We assume that only a part of available leading predictor variables should be used in forecasting various target variables. Hereafter two types of applied selection procedures are adduced.

2.3.1. Hard thresholding

Hard thresholding method uses statistical significance of a certain leading variable as a selection criterion. In other words, making a decision about inclusion of the variable in a targeted set we rely on the t-statistics of the coefficient δ in the following model:

$$y_{t+h} = \alpha + \sum_{j=0}^{r} z_{t-j\cdot}\beta_j + x_{t-h,i}\delta + \varepsilon_{t+h}\,. \qquad (15)$$

This approach is similar to the supervised principal components method of Bair, Hastie, Paul, and Tibshirani, R. (2006), but takes into account time series structure of the data.

As a result, in a targeted set we include only the leading variables with larger absolute values of the corresponding t-statistics. We use three different cut-off levels as in Bai and Ng (2008): the lowest acceptable absolute values of t-statistics are 1.28, 1.65 or 2.58.

2.3.2. Soft thresholding

The main pitfall of the hard thresholding is that leading variables are selected without consideration of other available predictors. Eventually we have a lot of highly correlated, or “similar”, variables in a targeted set. Soft thresholding method uses an alternative approach, which takes into account previously selected variables. An approach described hereunder is known as Least-angle regression (LARS) as in Efron, Hastie, Johnstone, and Tibshirani (2004).

Before applying LARS one should transform the data. The target variable should be centered, columns of the leading predictors matrix X should be centered and have a unit length. The algorithm consists of the following steps:

1. Set a forecast of the target variable as $\hat{Y}_0 = 0$.
2. Compute a vector of current correlations $\hat{c} = X'(Y - \hat{Y}_0)$.
3. Define an active set of indexes M corresponding to a maximum correlation value:

   $C = \max_j |\hat{c}_j| \qquad M = \{j : |\hat{c}_j| = C\}.$
4. Define an active matrix of predictors X_M:

   $X_M = (s_j x_{\cdot j})_{j \in M}$, where $s_j = sign(\hat{c}_j)$.
5. Compute a unit equiangular vector u_M - a vector equally correlated with all columns of the active matrix X_M:

   $u_M = X_M w_M$, where $w_M = B_M G_M^{-1} e_M$,

   $B_M = (e'_M G_M^{-1} e_M)^{-1/2}$,

   $G_M = X'_M X_M$ and

   e_M is a unit vector of size of the set M.
6. Define vector b as:

   $b = X' u_M$
7. Update the forecast of the target variable:

   $\hat{Y}_1 = \hat{Y}_0 + \hat{\theta} u_M$, where $\hat{\theta} = \min^{+}_{j \in M^c} \left(\frac{C - \hat{c}_j}{B_M - b_j}, \frac{C + \hat{c}_j}{B_M + b_j}\right)$ and minimum is taken over only positive components.

8. Go to step 2, increase all the subscripts of the forecasted value by one. Continue until all the indexes are in the active set.

At each loop one index is added to the active set. Therefore, upon completion of the algorithm we have a list of the leading predictor variables ordered as they were included into the active matrix. Let *m* be a dimension of the set M. A cut-off level of the targeted set of predictors is defined according to the smallest outcome of the SIC:

$$m^* = \arg\min_m SIC(m) = \ln(\text{var}(Y - \hat{Y}_M)) + m \ln T / T. \tag{16}$$

Chapter 3. Data description and forecasts evaluation

In this Chapter we discuss in more details the data used, targeted and leading variables, out-of-sample forecasting procedure and methods of comparing different forecasts.

3.1. Data description

The main part of the empirical application is based on the data from Stock and Watson (2005). The data consists of monthly observations on 128 U.S. macroeconomic variables over the period from January 1959 to December 2003, in total 540 observations for each variable. The series fall into 14 different categories, the categorization is summarized in Table 1.

Table 1. Categories of predictor variables

Category name	Number of series
Real output and income	15
Employment and hours	29
Real retail	1
Consumption	1
Housing starts and sales	10
Real inventories	3
Orders	7
Stock prices	4
Exchange rates	5
Interest rates and spreads	17
Money and credit quantity aggregates	11
Price indexes	21
Average hourly earnings	3
Consumer expectations	1

The variables are transformed into stationary series by taking logarithms and/or first differences. Generally, logarithms are used for housing starts and sales, first differences for nominal interest rates, first differences of logarithms for real quantity variables and employment, second differences of logarithms for price indexes, money and credit aggregates, and earnings. Details on the transformations as well as a complete overview of the variables are given in

Appendix A. For more detailed information on the variables we refer to Business Cycle Indicators Handbook (2001).

Target variables to be forecasted are *h*-months-ahead annualized growth rates of the following variables:

1. Composite coincident index (CCI) of The Conference Board;
2. Industrial production index;
3. Personal income less transfers (bil.dollars, chain 2000);
4. Total nonagricultural employment (thous.people);
5. Manufacturing and trade sales (mil.dollars, ahcin 1996).

A growth rate of a variable s_t over h months is computed as:

$$y_{t+h} = \frac{1200}{h} * \ln \frac{s_{t+h}}{s_t}. \tag{17}$$

Four last target variables are entries for computing Composite coincident index. A growth rate of the target variable over a previous month is used as a preferential predictor.

3.2. Other benchmark models

We compare not only forecasts produced by different factor models, by as well by more common benchmarks. As base models we consider univariate autoregression and autoregression with distributed lags.

Univariate autoregressive forecasting is based on a model which allows for direct *h*-month-ahead forecasts, with the following forecast equation:

$$y_{t+h} = \alpha_0 + \sum_{j=1}^{r} y_{t-j+1}\alpha_j + \varepsilon_{t+h}. \tag{18}$$

The autoregressive lag order r is chosen according to the smallest outcome of the SIC.

ADL forecasting is based on the same model as DFM (10). But instead of the factor Composite leading index[10] of the Conference board is used:

[10] CLI is defined as a weighted average of 10 macroeconomic indicators. The list is presented in Appendix A. For more details see Business Cycle Indicators Handbook (2001)

$$y_{t+h} = \alpha + \sum_{j=0}^{r} z_{t-j}.\beta_j + \sum_{j=0}^{q} CLI_{t-j}\gamma_j + \varepsilon_{t+h}\,. \tag{19}$$

The lag orders r and q are chosen according to the smallest outcome of the SIC.

3.3. Stepwise forecasting

In this research we evaluate forecasting power of different models basing on out-of-sample forecasts. Moving window method is applied. Roughly speaking, if w is a window width in terms on a number of observations, then we use all the information available over a period $[t_0 - w, t_0 - 1]$ for model estimation. Inserting estimations into forecasting equation we obtain a h-months-ahead forecast $\hat{y}_{t_0+h}$ at the moment t_0.At the next step we use data over a period $[t_0 + 1 - w, t_0]$ to forecast at the moment $t_0 + 1$, and so on. Since our models require standardization of the data inside a window and use of lagged values, in practical work slightly different windows are used. The approache applied is presented in Heij, Groenen, and Van Dijk (2008). The moment of the first forecast depends on the first available observation of the series and a window width. The moment of the last forecast depends on the last available observation and a forecasting horizon. We consider 1-, 3-, 6- and 12-months-ahead forecasts. As a result we have a series of forecasted values, forecasting accuracy is evaluated by forecasting errors $e_t = y_{t+h} - \hat{y}_{t+h}$ analysis.

3.4. Forecasts evaluation

The main tool used to compare different forecasts of the same target variables over the same forecasting horizon is the mean squared prediction error:

$$MPSE = \frac{1}{T - h - t_o + 1} \sum_{t_o}^{T-h} e_t^2\ . \tag{20}$$

We use MSPE over the period from 1960 to 2003 for comparative analysis of the competing models in different specifications. Later on we analyze ability of the most successful models to forecast a cyclical phase over the period from January 2004 to August 2009. The data over that period is not compete and has some missing values, and so we use the data only till 2003 for the main part of the research.

Cyclical phase forecasting is based on the growth rates forecasts of CCI. Recession is defined as a negative growth over two subsequent quarters. In case growth signs alternate, the phase is called mixed. Hence recession indicator over the following two quarters is defined as:

$R_t = 1$, if $y_{t+3} > 0$ and $y_{(t+3)+3} < 0$;

$R_t = 0$, if $y_{t+3} > 0$ and $y_{(t+3)+3} > 0$;

$R_t = 0{,}5$, otherwise.

Forecasted values $\hat{y}_{t+3}$ and $\hat{y}_{t+6}$ are converted into forecasted probability of recession. Since $\hat{y}_{t+3}$ and $\hat{y}_{t+6}$ are annualized, we firstly convert them into quarterly growth rates:

$$\hat{y}_{1Q,t} = 1/4 * \hat{y}_{t+3}, \tag{21}$$

$$\hat{y}_{2Q,t} = 1/2 * \hat{y}_{t+6} - 1/4 * \hat{y}_{t+3}, \tag{22}$$

where $\hat{y}_{1Q,t}$ is an estimate of y_{t+3}, $\hat{y}_{2Q,t}$ is an estimate of $y_{(t+3)+3}$.

Recession probability over the following two quarters is estimated as:

$$\hat{p}_t = \hat{\Pr}(y_{1Q,t} < 0 \cup y_{2Q,t} < 0), \tag{23}$$

assuming that $y_{1Q,t}$ and $y_{2Q,t}$ are jointly normally distributed with the mean vector $[\hat{y}_{1Q,t}, \hat{y}_{2Q,t}]$ and the covariance matrix estimated over the last 120 actual observations. Probability $\hat{p}_t$ is converted into a binary signal $\hat{R}_t$, which takes a value of one if $\hat{p}_t$ is larger than average R_t over the last 120 observations and a value of zero otherwise.

Chapter 4. Results

4.1. Variables selection results

Firstly, we would like to present an analysis of variables selected as targeted predictors. In Table 2 selection results for CCI as a target variable are summarized. Since selection procedures were performed in every window we are able to calculate the following statistics:

1. Average number of variables selected as targeted (out of 128);
2. Number of variables selected with frequency 80% and more;
3. Number of variables selected with frequency 20% and less;
4. 10 most frequently selected variables. In the Table only mnemonics are presented, full description is presented in Appendix A.

Presented figures show difference between hard and soft thresholding in empirical application: the soft one selects much less variables generally from different categories. Variables from the same category are often mutually correlated. Hard classifier analyses significance of each variable one by one and usually selects all correlated group. But soft classifier, if one of the correlated variables is already selected, usually skips others from the correlated group. Hard classifier, even under the less tight threshold of 1.28, cuts off more than a half of variables as irrelevant. Soft classifier selects typically not more than 10 variables as targeted, even so none of them is selected with frequency 80% and more.

Further we consider the most frequently selected variables. Generally top-ten variables are similar for short and long forecasting horizons. Housing starts (hsfr), housing authorized (hsbr), ratio of help-wanted advertising to a number of unemployed (lhelx) are selected with frequency more than 80 % for all horizons. But differences are also observed. New orders index (pmno) and production index (pmp) are important indicators for short- (1 month) and middle-term (3 and 6 months), but not for long-term (12 months) forecasting of the CCI growth rate. For the short forecasting horizon variables from the categories “Housing starts and sales”, “Orders” and “Employment and hours” are important. The most typical variables are purchasing managers’ index (pmi), and a number of employees in different sectors (ces003, ces015, ces033). For the middle horizons variables from the categories “Employment and hours” and “Orders” (excluding pmno) become less important.

Table 2. Variables selection statistics, CCI growth rates forecasting

	1m	3m	6m	12m	1m	3m	6m	12m
	Hard thresholding (1.28)				Hard thresholding (1.65)			
Average number of selected variables	55,31	51,40	49,32	50,47	41,03	38,91	39,90	39,95
Number of variables selected with frequency 80% and more	18	18	23	23	10	12	18	18
Number of variables selected with frequency 20% and less	31	35	59	52	53	68	69	68
10 most frequently selected variables	pmno*	pmno	pmno*	lhel	pmi	pmno	hsfr	sfyaaac
	pmi	hsbr	hsfr	fm2dq	pmno	hsbr	pmno	sfygt10
	pmp	lhelx	hsbr	fsdxp	pmp	hsfr	hssou	Lhelx
	ces003	hsfr	hssou	hsfr	ces003	hssou	hsbr	sfybaac
	hsfr	hssou	pmp	lhelx	hsfr	lhelx	fm2dq	fm2dq
	hsbr	pmp	sfyaaac	sfyaaac	hsbr	pmp	lhelx	sfygt5
	hssou	fm2dq	fm2dq	sfygt10	ces015	fm2dq	sfygt10	Hsbr
	ces033	lhel	lhelx	fspcom	lhelx	sfyaaac	sfyaaac	Lhel
	lhelx	sfyaaac	sfygt10	fspin	Hssou	sfybaac	pmp	Pmdel
	sfygt1	sfygt10	fspin	sfybaac	lhel	sfygt10	sfygt5	Fsdxp
	Hard thresholding (2.58)				Soft thresholding			
Average number of selected variables	18,19	20,37	24,50	24,77	4,18	9,06	11,72	11,29
Number of variables selected with frequency 80% and more	2	3	7	5	0	0	0	0
Number of variables selected with frequency 20% and less	95	92	86	85	123	113	105	103
10 most frequently selected variables	pmno	hsfr	hsfr	sfyaaac	pmno	pmno	pmno	fm2dq
	pmp	hsbr	hsbr	sfygt10	lhel	hsbr	hsbr	sfyaaac
	hsbr	pmno	pmno	sfybaac	ces033	lhel	ces033	fclbmc
	hsfr	pmp	sfyaaac	sfygt5	ips10	ces033	lhel	Hsbr
	lhel	hssou	sfybaac	fm2dq	sfygm3	lhelx	Fm2dq	Pmno
	hswst	fm2dq	hssou	sfygm6	ips299	sfygm6	sfygm6	Pmcp
	pmi	sfyaaac	lhelx	sfygm3	a1m092	hssou	fybaac	Sfygm6
	sfygm6	lhelx	pmp	lhel	pmi	a1m092	fclbmc	sfygt10
	sfygm3	sfygt10	fm2dq	lhelx	hsbr	sfyaaac	hssou	hssou
	ces015	sfygm6	sfygt10	pmdel	hsfr	fclbmc	pmcp	sfybaac

Variables selected with frequency 100% are marked with asterisk (*)

But more frequently variables from the categories “Interest rates and spreads” and “Money and credit quantity aggregates” are selected, particularly money supply M2 (fm2dq), federal funds interest rate and AAA corporate bond yield spread (sfyaaac), federal funds interest rate and 10-years treasury interest rate spread (sfygt10). For the long forecasting horizon variables from the categories "Interest rates and spreads" and "Money and credit quantity aggregates" are the most important ones, under the threshold of 1,28 "Stock prices" are also selected quite often. Thus we observe that for short-term economic growth rate forecasting one should watch over the dynamics of new housing starts and orders. For longer horizons money supply and interest rates' spreads become more important. This conclusion corresponds with other findings that interest rates have a longer lead time than other indicators.

For other four target variables we present only key features. Tables similar to Table 2 could be found in Appendix B.

For industrial production index it is typical that short-term treasury and federal funds interest rates spreads and important for all horizons. In general we observe spreads in top-ten for all horizon and all classifiers. However only for the long horizon spreads dominate. New orders index (pmno) is important for short- and middle-term forecasting. Production index (pmp) is frequently selected for 1- and 3-months-ahead forecasting. For the short horizon variables from the categories "Housing starts and sales" and "Employment and hours" are also important. Noteworthily index of consumer expectations (hhsntn) is important for middle-term dynamics. Price indexes (fsdxp, fspin) and corporate bonds spreads (sfyaaac, sfybaac) are more frequently selected for longer horizons.

Among variables selected for forecasting personal income growth rates there are no variables systematically important for all forecasting horizons. For the long horizon again interest rates spreads dominate, as well as index of help-wanted advertising (lhel) and ratio of this index to a number of unemployed (lhelx). For the short and middle horizons new orders index and a number of employees in different sectors are very important. For middle-term forecasting importance of a number of employees decreases, but we observe more variables, related to dynamics of real production index. Soft thresholding results are different: production index is important for short-term forecasting, spreads are not so important for longer horizons.

When forecasting employment growth rates, it is hard to mark out variables specific for one or another horizon. For all horizons interest rates spreads, new orders index and "Employment and hours" category are important. For shorter horizons category "Housing starts and sales" is specific. For longer ones – "Stock prices" category. Soft classifier marks variables from the category "Orders" as targeted predictors, but spreads become less important.

For manufacturing and trade sales interest rates' spreads and new orders index dominate for all forecasting horizons. For shorter horizons personal income and employment are also systematically selected. Soft thresholding results are considerably different. Variables from the category "Money and credit quantity aggregates" dominate in the middle and long horizons, in the short one spreads again become less important.

Summing up we could state that variables selection results in general do correspond to our expectancies. More than a half of available variables were marked as irrelevant. Interest rates variables are systematic targeted predictors for long-term growth rates forecasting. Housing starts dynamics is more important for short-term forecasting. The usual suspects were observed in the top-ten list for specific target variables: consumer expectations for industrial production, employment for personal income. Such conclusions allow us to expect increased forecasting accuracy of factor models augmented by a preliminary step of targeted variables selection.

4.2. Growth rates forecasting results

Mean squared prediction errors are presented in Table 3 and in Tables 8-11 in Appendix B. MSPE's are given relatively to the variance of the corresponding target variable. Forecast accuracy was analyzed for the whole sample and for three subsamples – from 1970 to 1983, from 1984 to 1993 and from 1994 to 2003. The lowest MSPE for the following groups are given in bold font:

- Benchmark models (univariate autoregression and autoregression with distributed lags);
- Principal components regression models;
- Principal covariate regression model.

The lowest MSPE for specific horizon and subsample are underlined. Firstly we discuss forecasting power of different models for each target variables and then present general conclusions.

4.2.1. Composite coincident index

For all forecasting horizons and subsamples, except 6 and 12 months over the period from 1994 to 2003, factor models are more accurate. Reduction of the MSPE is 22% on average. The most substantial gain of factor models is observed over the most volatile period from 1970 to 1983. In general targeted predictors estimate the factor with higher predictive power for the middle and long forecasting horizons. Only hard thresholding classifier is effective. Gain of combining variables selection and principal covariate regression is usually smaller than of combining variables selection and principal component regression. Even so, precisely principal covariate regression models with hard classifiers allows for the most accurate middle- and long-term forecasts. Principal covariate regression models are relatively more effective in short-term forecasting and for the periods of less volatile dynamics of the target variable.

4.2.2. Industrial production index

For industrial production index growth rates over the period from 1984 to 1993 are the most stable in terms of variance and relatively accurate forecasts are obtained using composite coincident index as a factor (autoregression with distributed lags) for all forecasting horizon. In other cases

Table 3. Mean squared prediction errors relative to variance, CCI growth rates forecasting

				Principal component regression					Principal covariate regression				
Sample	Variance	AR	CLI	no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)	no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)
1 month													
1970-2003	18,181	0,864	**0,836**	0,748	0,782	**0,719**	0,722	0,772	0,733	0,784	**0,719**	0,761	0,762
1970-1983	30,027	0,790	**0,748**	0,671	0,687	0,643	**0,626**	0,684	**0,651**	0,689	0,654	0,688	0,652
1984-1993	11,416	1,069	**1,065**	0,976	0,957	**0,912**	0,985	0,994	1,014	0,957	**0,944**	0,976	1,022
1994-2003	8,263	**0,958**	0,971	**0,831**	1,029	0,843	0,851	0,918	0,762	1,031	**0,743**	0,840	0,965
3 months													
1970-2003	10,657	0,802	**0,707**	0,666	0,622	0,606	**0,583**	0,595	0,605	0,597	0,579	0,591	**0,562**
1970-1983	19,425	0,808	**0,692**	0,659	0,564	0,572	0,541	**0,537**	0,578	0,521	0,540	0,545	**0,486**
1984-1993	4,258	0,925	**0,857**	**0,722**	0,871	0,764	0,757	0,812	0,818	0,913	**0,783**	0,836	0,866
1994-2003	4,617	**0,651**	0,654	**0,654**	0,740	0,668	0,673	0,741	**0,568**	0,758	0,618	0,639	0,733
6 months													
1970-2003	8,167	0,908	**0,652**	0,780	0,749	**0,541**	0,569	0,564	0,610	0,635	0,564	**0,528**	0,549
1970-1983	14,670	0,962	**0,626**	0,797	0,693	**0,458**	0,487	0,470	0,555	0,509	0,471	0,439	**0,431**
1984-1993	2,995	**0,919**	0,931	**0,765**	0,980	0,776	0,802	0,860	0,977	1,148	0,966	**0,882**	1,056
1994-2003	4,005	0,613	**0,582**	**0,703**	0,873	0,807	0,829	0,844	**0,627**	0,919	0,753	0,729	0,788
12 months													
1970-2003	5,962	0,974	**0,757**	0,957	0,657	0,655	0,605	**0,581**	0,672	0,686	0,535	0,547	**0,523**
1970-1983	10,232	1,034	**0,738**	1,018	0,537	0,561	0,461	**0,435**	0,531	0,497	0,376	0,371	**0,335**
1984-1993	2,159	**0,845**	0,936	**0,730**	0,923	0,932	1,038	0,979	1,422	1,332	1,195	1,176	**1,078**
1994-2003	3,477	0,806	**0,736**	**0,854**	1,038	0,905	0,976	0,987	**0,816**	1,120	0,820	0,931	1,010

factor models result with substantially lower MSPE, with reduction by 23 % on average. Soft classifier is again less effective than hard thresholding. In general principal covariate regression models allows for more accurate forecasts. Over more volatile periods use of hard classifier is worthwhile, over the less volatile period from 1993 to 2003 no selection procedure is needed for 1- and 3-months ahead forecasting.

4.2.3. Personal income

In personal income growth rates forecasting, factor models systematically outperform benchmarks. Results for the short- and middle-term horizons are very similar. For these horizons dominance of principal covariate regression in combination with hard thresholding is obvious. However, pairwise comparison of factor models indicates that selection procedures are generally not effective for short-term forecasting. For the long horizon principal component regressions are the most accurate models for all subsamples. Characteristic feature of the target variable is that relative forecasting accuracy of the models does not depends on volatility over a given subsample.

4.2.4. Nonagricultural employment

Over the periods from 1984 to 1993 and from 1994 to 2004 dramatic reduction of growth rates variance is observed. Over these periods factor models are not able to outperform autoregression and autoregression with distributed lags. Over the periods with relatively high volatility of the target variables principal component regression dominates on the short horizon, principal covariate regression dominates on the middle and long horizons. Use of factor models reduces the mean square prediction error by 22 % on average, for longer horizons reduction is more substantial. Forecasting accuracy is increased by variables selection, but more considerable gain is observed for longer forecasting horizons.

4.2.5. Manufacturing and trade sales

Manufacturing and trade sales growth rates are the most volatile over the examined ones, they are usually difficult for forecast. Although factor models dominated for 1- and 3-months ahead forecasting, MSPE reduction is only 12% at maximum. For longer horizons gain in much more substantial, reduction is 27% on average. Dominance of the benchmark models over the less volatile periods is observed only for middle-term forecasting. Use of principal covariate regression is reasonable only for 3- and 12-months-ahead forecasting. Variables selection gain increases with forecasting horizon.

4.2.6. General results

The main conclusions of relative forecasting accuracy of compared models could be summarized as follows:

- Factor models reduce the mean squared prediction error by 20 % on average as compared to the benchmark models.
- Factor models are relatively more accurate in forecasting over the periods of higher volatility of the target variable.
- Factor estimated from a set of targeted predictors has higher predictive power.
- Variables selection gain increases with forecasting horizon. Consequently, for longer forecasting horizons maximum MSPE reduction in factor models is observed.
- Except some cases, principal covariate regression model dominates. Exceptions are long-term forecasting of personal income and short-term forecasting of industrial production, employment and sales.

Cyclical phase forecasting results

We present below the plots of observed recession indicator R_t, recession indicator as in National Bureau of Economic Research (NBER) report, and binary variable $\hat{R}_t$ estimates based on the following forecasting models:

- Autoregression with distributed lags, as in (19);
- Factor models without variables selection:
  - Principal component regression, as in (9) and (10);
  - Principal covariate regression, as in (11);
- Factor models with hard thresholding classifier, cut-off level is 2.58:
  - Principal component regression, as in (9) and (10);
  - Principal covariate regression, as in (11);

According to the NBER report, recent recession began in December 2007[11]. Observed recession indicator R_t was calculated on the basis of dynamics direction (positive or negative) of the Composite coincident index. NBER's Business Cycle Dating Committee makes decisions on recession dating basing on more information than one index and over more than a six-month period.

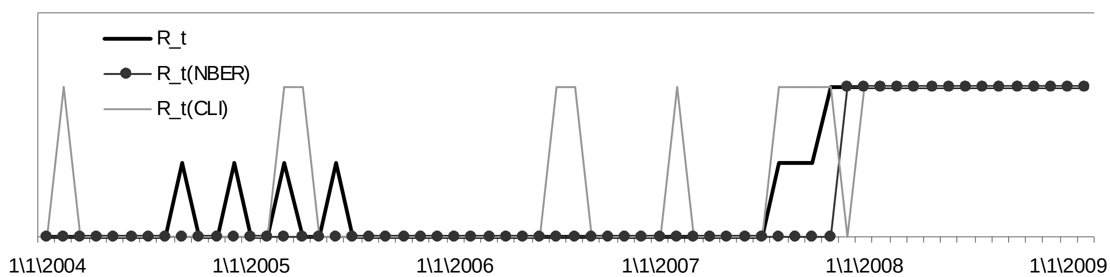

Figure 1. Recession forecasting: autoregression with distributed lags.

[11] See http://www.nber.org/cycles/dec2008.html for more details.

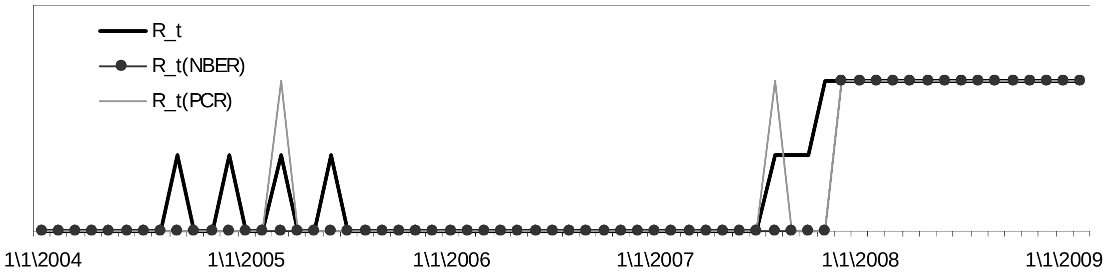

Figure 2. Recession forecasting: principal component regression.

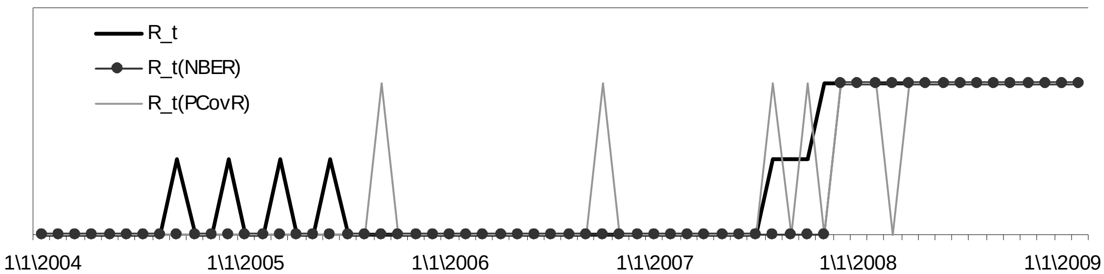

Рисунок 3. Recession forecasting: principal covariate regression.

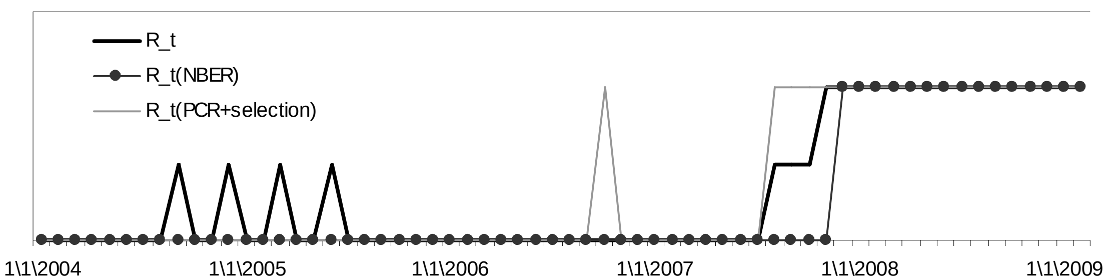

Figure 4. Recession forecasting: predictor variables selection and principal component regression.

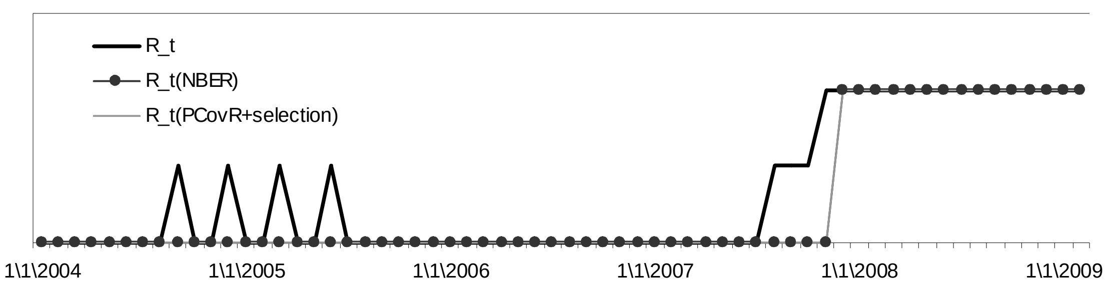

Figure 5.Recession forecasting: predictor variables selection and principal covariate regression.

Even the most simple autoregression model with distributed lags was able to indicate the recent recession without substantial delay. But on the other hand a lot of “false” recession signals are observed. Factor models without predictor variables selection are less precise in recession dating and forecasts a lot of “false” recessions in the second half of 2007. “False” signals of economics downturn are also observed in 2005 and 2006. Factor models with predictor variables selection are the most accurate in recession dating. Principal component regression forecast coincides with the computed recession indicator R_t, but one "false" signal is observed at the turn of 2006 and 2007 years. Principal covariate regression forecast of the binary variable $\hat{R}_t$ completely concurs with the NBER recession indicator without any "false" signals.

Therefore, preliminary selection of targeted variables helps to get rid of "noisy" and irrelevant predictors and reduce a number of "false" recession signals. Both considered factor estimation methods (PCR and PCovR) ends up with similar and highly accurate forecasts of economic cyclical phases.

Chapter 5. Conclusions

In forecasting future dynamics of the economy in general or any specific economic indicator a researcher faces two key problems: which model to use as a base one - parsimonious and theoretically-founded or more sophisticated statistically-econometric; which variables to take as exogenous.

Lately more and more researchers and decision makers make use of information and technological progress, notably of prompt availability of high-frequency data over a huge amount of economic variables. Use of complicated econometric models employing statistical properties of virtually unlimited number of exogenous variables is already far beyond just pure academic papers. The presented work considers dynamic factor models (DFM) for a forecasting purpose. This type of models is currently used by public authorities in the U.S. and European Union.

Factor models hypothesize that a big group of observed economic variables vary with time under the influence of a limited number of common trends (factors) and individual idiosyncratic shocks. We consider to approaches to the factors estimation: (1) principal *component* regression; (2) principal *covariate* regression as in Heij et al. (2006). Principal covariate regression explicitly takes into account the forecasting power of the extracted factors.

Relatively small number of researchers has investigated a question of selection of exogenous variables, which are used in the factors estimation. Bai and Ng (2008) showed that including irrelevant predictors into a model could substantially lower forecasting accuracy. In our research we compared features of principal component and covariate regressions and analyzed an impact of preliminary variables selection on forecasting accuracy.

In the empirical part U.S. monthly data on 128 macroeconomic variables running from 1960 were analyzed. As target variables we used: the composite coincident index, the industrial production index, personal income less transfers, total nonagricultural employment, manufacturing and trade sales. We applied a moving-window approach in order to obtain out-of-sample forecasted values. Mean squared prediction error was used as a criterion of the forecasting accuracy.

We have discovered that in the considered settings use of dynamic factors for forecasting is often proved only for the periods of relatively high volatility of the target variable. The inquiry is that in the beginning of 1980th a substantial domestic change of the economy occurred in the U.S., more known as the Great Moderation. At this moment volatility of many economic variables reduced more that twice. There is no a common opinion about reasons of such an abrupt change. Theoretical explanations vary from an increased reasoning behind macroeconomic policy to a coincidence. Howbeit, over the period after structural change forecasting accuracy of factor models is considerably lower than of autoregression and autoreagression with distributed lags. DFM's failure is explained by their exploitation of the data over 10 years preceding a forecasting moment. In this case a lot of observations are taken from the period before the structural change. But benchmark models take into account not more than preceding year and a half. In fact, after 1994 we again observed a general increase of relative forecasting accuracy of DFM. And, last but not least, dynamic factor models were hardly ever outperformed on long forecasting horizons.

In addition we have empirically supported the hypothesis that factors extracted from a pre-selected set of targeted variables have a higher predictive power. We considered two alternative approaches of variables selection: hard thresholding and soft thresholding. Soft classifier occurred to be ineffective for the investigated problem. Soft classifier tends to select a sparser set of targeted variables. But a set of a small number of uncorrelated variables is usually an inappropriate base for principal components extraction.

Forecasted values of composite coincident index growth rates were used for estimation of recession probability. Comparison of various models displayed that factor models provide much more precise forecasts of economic cyclical phases. And exogenous variables selection allows to get rid of “noisy” and irrelevant predictors and reduce a number of “false” recession signals.

Therefore we conclude:

- Dynamic factors are important indicators of the economic variables’ dynamics.
- One has to conduct preliminary examination of available exogenous variables for their relevance in forecasting a certain target variable.
- Principal covariate regression has a built-in procedure of variables selection, so considered explicit methods of hard and soft thresholding are not as effective as for principal component regression.
- Combined use of predictor variables selection and principal covariate regression gives a greater effect in forecasting accuracy increase.

Bibliography

1. Bai, J., Ng, S. (2002). “Determining the Number of Factors in Approximate Factor Models”. *Econometrica*, 70, 191-221.
2. Bai, J., Ng, S. (2006). “Evaluating Latent and Observed Factors in Macroeconomics and Finance”. *Journal of Econometrics*, 131, 507-537.
3. Bai, J., Ng, S. (2008). “Forecasting Economic Time Series Using Targeted Predictors”. *Journal of Econometrics*, 146, 304-317.
4. Bair, E., Hastie, T., Paul, D., Tibshirani, R. (2006). “Prediction by Supervised Principal Components”. *Journal of the American Statistical Association*, 101(473), 119-137.
5. Bernanke, B.S., Boivin, J., Eliasz, P. (2005). “Measuring the Effectsof Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach”. *Quarterly Journal of Economics*, 120, 525-546.
6. Boivin, J., Ng, S. (2006). “Are More Data Always Better for Factor Analysis?” *Journal of Econometrics*, 132, 169-194.
7. Burns, A.F., Mitchell, W.C. (1946). “Measuring Business Cycles”. *NBER Studies in Business Cycles* No. 2,New York.
8. Conference Board, The (2001). Business Cycle Indicators Handbook.
9. Croux, C., Renault, E., Werker, B. (2004). “Dynamic Factor Models”. *Journal of Econometrics*, 119, 223-230.
10. De Jong, S., Kiers, H.A.L. (1992). “Principal covariate regression”. *Chemometrics and Intelligent Laboratory Systems*, 14, 155-164.
11. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). “Least Angle Regression”. *The Annals of Statistics*, 32(2), 407-499.
12. Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2000). “The Generalized Dynamic Factor Model: Identification and Estimation”. *Review of Economics and Statistics*, 82, 540-554.
13. Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2003). “Do Financial Variables Help Forecasting Inflation and Real Activity in Euro Area”. *Journal of Monetary Economics*, 50, 1243-1255.

14. Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2004). “The Generalized Dynamic Factor Model: Consistency and rates”. *Journal of Econometrics*, 119, 231-255.
15. Gelper, S., Croux, C., (2008), "Time series least angle regression for selecting predictive economic sentiment series". *Working Paper.*
16. Heij, C., Groenen, P.J.F., D. van Dijk (2006). “Time Series Forecasting by Principal Covariate Regression”. *Econometric Institute Report EI2006-37*, Rotterdam.
17. Heij, C., Groenen, P.J.F., D. van Dijk (2007). “Forecast Comparison of Principal Component Regression and Principal Covariate Regression”. *Computational Statistics and Data Analysis*, 51, 3612-3625.
18. Heij, C., Groenen, P.J.F., D. van Dijk (2008). “Macroeconomic Forecasting with Matched Principal Components”. *International Journal of Forecasting*, 24, 87-100.
19. Keirs, H.A.L. (1990). “Majorization as a tool for optimizing a class of matrix functions”. Psychometrika 55, 417-428.
20. Marcellino, M.G. (2006). “Leading Indicators”. *Handbook of Economic Forecasting*, Volume 1, edited by Graham Elliott, Clive W.J. Granger and Allan Timmermann, 879-960.
21. Moench, E. (2008). “Forecasting the Yield Curve in Data-Rich Environment: a No-Arbitrage Factor-Augmented VAR Approach”. *Journal of Econometrics*, 99, 39-61.
22. Stock, J.H., Watson, M.W. (2002). “Macroeconomic Forecasting Using Diffusion Indexes”. *Journal of Business and Economic Statistics*, 20, 147-162.
23. Stock, J.H., Watson, M.W. (2005). “An Empirical Comparison of Methods for Forecasting Using Many Predictors”. *Working Paper.*
24. Stock, J.H., Watson, M.W. (2006). “Forecasting with Many Predictors”. *Handbook of Economic Forecasting*, Volume 1, edited by Graham Elliott, Clive W.J. Granger and Allan Timmermann, 515-554.

25. Zou, H., Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net". *Journal of Royal Statistical Society*, Series B 67(2), 301-320.

26. http://www.conference-board.org/

27. http://www.chicagofed.org/webpages/research/data/cfnai/current_data.cfm

28. http://eurocoin.cepr.org/

Appendix A

List of the exogenous variables: mnemonics, transformation type, variable description and category.

Mnemonic	Transf.	Description	Category
a0m052	Δln	Personal Income (AR, Bil. Chain 2000 $) (TCB)	Real Output and Income
a0m051	Δln	Personal Income Less Transfer Payments (AR, Bil. Chain 2000 $) (TCB)	Real Output and Income
a0m224_r	Δln	Real Consumption (AC) a0m224/gmdc (a0m224 is from TCB)	Consumption
a0m057	Δln	Manufacturing And Trade Sales (Mil. Chain 1996 $) (TCB)	Manufacturing and Trade Sales
a0m059	Δln	Sales Of Retail Stores (Mil. Chain 2000 $) (TCB)	Real Retail
ips10	Δln	Industrial Production Index - Total Index	Real Output and Income
ips11	Δln	Industrial Production Index - Products, Total	Real Output and Income
ips299	Δln	Industrial Production Index - Final Products	Real Output and Income
ips12	Δln	Industrial Production Index - Consumer Goods	Real Output and Income
ips13	Δln	Industrial Production Index - Durable Consumer Goods	Real Output and Income
ips18	Δln	Industrial Production Index - Nondurable Consumer Goods	Real Output and Income
ips25	Δln	Industrial Production Index - Business Equipment	Real Output and Income
ips32	Δln	Industrial Production Index – Materials	Real Output and Income
ips34	Δln	Industrial Production Index - Durable Goods Materials	Real Output and Income
ips38	Δln	Industrial Production Index - Nondurable Goods Materials	Real Output and Income
ips43	Δln	Industrial Production Index - Manufacturing (Sic)	Real Output and Income
ips307	Δln	Industrial Production Index - Residential Utilities	Real Output and Income
ips306	Δln	Industrial Production Index – Fuels	Real Output and Income
pmp	lv	Napm Production Index (Percent)	Real Output and Income
a0m082	Δlv	Capacity Utilization (Mfg) (TCB)	Real Output and Income
lhel	Δlv	Index Of Help-Wanted Advertising In Newspapers (1967=100;Sa)	Employment and Hours
lhelx	Δlv	Employment: Ratio; Help-Wanted Ads:No. Unemployed Clf	Employment and Hours
lhem	Δln	Civilian Labor Force: Employed, Total (Thous.,Sa)	Employment and Hours
lhnag	Δln	Civilian Labor Force: Employed, Nonagric.Industries (Thous.,Sa)	Employment and Hours
lhur	Δlv	Unemployment Rate: All Workers, 16 Years & Over (%,Sa)	Employment and Hours
lhu680	Δlv	Unemploy.By Duration: Average(Mean)Duration In Weeks (Sa)	Employment and Hours
lhu5	Δln	Unemploy.By Duration: Persons Unempl.Less Than 5 Wks (Thous.,Sa)	Employment and Hours
lhu14	Δln	Unemploy.By Duration: Persons Unempl.5 To 14 Wks (Thous.,Sa)	Employment and Hours
lhu15	Δln	Unemploy.By Duration: Persons Unempl.15 Wks + (Thous.,Sa)	Employment and Hours
lhu26	Δln	Unemploy.By Duration: Persons Unempl.15 To 26 Wks (Thous.,Sa)	Employment and Hours
lhu27	Δln	Unemploy.By Duration: Persons Unempl.27 Wks + (Thous,Sa)	Employment and Hours
a0m005	Δln	Average Weekly Initial Claims, Unemploy. Insurance (Thous.) (TCB)	Employment and Hours
ces002	Δln	Employees On Nonfarm Payrolls: Total Private	Employment and Hours
ces003	Δln	Employees On Nonfarm Payrolls - Goods-Producing	Employment and Hours
ces006	Δln	Employees On Nonfarm Payrolls – Mining	Employment and Hours

Mnemonic	Transf.	Description	Category
ces011	Δln	Employees On Nonfarm Payrolls – Construction	Employment and Hours
ces015	Δln	Employees On Nonfarm Payrolls - Manufacturing	Employment and Hours
ces017	Δln	Employees On Nonfarm Payrolls - Durable Goods	Employment and Hours
ces033	Δln	Employees On Nonfarm Payrolls - Nondurable Goods	Employment and Hours
ces046	Δln	Employees On Nonfarm Payrolls - Service-Providing	Employment and Hours
ces048	Δln	Employees On Nonfarm Payrolls - Trade, Transportation, And Utilities	Employment and Hours
ces049	Δln	Employees On Nonfarm Payrolls - Wholesale Trade	Employment and Hours
ces053	Δln	Employees On Nonfarm Payrolls - Retail Trade	Employment and Hours
ces088	Δln	Employees On Nonfarm Payrolls - Financial Activities	Employment and Hours
ces140	Δln	Employees On Nonfarm Payrolls – Government	Employment and Hours
a0m048	Δln	Employee Hours In Nonag. Establishments (AR, Bil. Hours) (TCB)	Employment and Hours
ces151	lv	Avg Weekly Hrs of Prod or Nonsup Workers On Private Nonfarm Payrolls - Goods-Producing	Employment and Hours
ces155	Δlv	Avg Weekly Hrs of Prod or Nonsup Workers On Private Nonfarm Payrolls - Mfg Overtime Hours	Employment and Hours
aom001	lv	Average Weekly Hours, Mfg. (Hours) (TCB)	Employment and Hours
pmemp	lv	Napm Employment Index (Percent)	Employment and Hours
hsfr	ln	Housing Starts:Nonfarm(1947-58);Total Farm&Nonfarm(1959-)(Thous.,Saar)	Housing Starts and Sales
hsne	ln	Housing Starts:Northeast (Thous.U.)S.A.	Housing Starts and Sales
hsmw	ln	Housing Starts:Midwest(Thous.U.)S.A.	Housing Starts and Sales
hssou	ln	Housing Starts:South (Thous.U.)S.A.	Housing Starts and Sales
hswst	ln	Housing Starts:West (Thous.U.)S.A.	Housing Starts and Sales
hsbr	ln	Housing Authorized: Total New Priv Housing Units (Thous.,Saar)	Housing Starts and Sales
pmi	lv	Purchasing Managers' Index (Sa)	Orders
pmno	lv	Napm New Orders Index (Percent)	Orders
pmdel	lv	Napm Vendor Deliveries Index (Percent)	Orders
pmnv	lv	Napm Inventories Index (Percent)	Real Inventories
a0m008	Δln	Mfrs' New Orders, Consumer Goods And Materials (Bil. Chain 1982 $) (TCB)	Orders
a0m007	Δln	Mfrs' New Orders, Durable Goods Industries (Bil. Chain 2000 $) (TCB)	Orders
a0m027	Δln	Mfrs' New Orders, Nondefense Capital Goods (Mil. Chain 1982 $) (TCB)	Orders
a1m092	Δln	Mfrs' Unfilled Orders, Durable Goods Indus. (Bil. Chain 2000 $) (TCB)	Orders
a0m070	Δln	Manufacturing And Trade Inventories (Bil. Chain 2000 $) (TCB)	Real Inventories
a0m077	Δlv	Ratio, Mfg. And Trade Inventories To Sales (Based On Chain 2000 $) (TCB)	Real Inventories
fm1	Δ^2ln	Money Stock: M1(Curr,Trav.Cks,Dem Dep,Other Ck'able Dep)(Bil$,Sa)	Money and Credit Quantity Aggregates
fm2	Δ^2ln	Money Stock:M2(M1+O'nite Rps,Euro$,G/P&B/D Mmmfs&Sav&Sm Time Dep(Bil$,Sa)	Money and Credit Quantity Aggregates
fm3	Δ^2ln	Money Stock: M3(M2+Lg Time Dep,Term Rp's&Inst Only Mmmfs)(Bil$,Sa)	Money and Credit Quantity Aggregates
fm2dq	Δln	Money Supply - M2 In 1996 Dollars (Bci)	Money and Credit Quantity Aggregates
fmfba	Δ^2ln	Monetary Base, Adj For Reserve Requirement Changes(Mil$,Sa)	Money and Credit Quantity Aggregates
fmrra	Δ^2ln	Depository Inst Reserves:Total, Adj For Reserve Req Chgs(Mil$,Sa)	Money and Credit Quantity Aggregates
fmrnba	Δ^2ln	Depository Inst Reserves:Nonborrowed,Adj Res Req Chgs(Mil$,Sa)	Money and Credit Quantity Aggregates
fclnq	Δ^2ln	Commercial & Industrial Loans Oustanding In 1996	Money and Credit

Mnemonic	Transf.	Description	Category
		Dollars (Bci)	Quantity Aggregates
fclbmc	lv	Wkly Rp Lg Com'l Banks:Net Change Com'l & Indus Loans(Bil$,Saar)	Money and Credit Quantity Aggregates
ccinrv	Δ^2ln	Consumer Credit Outstanding - Nonrevolving(G19)	Money and Credit Quantity Aggregates
a0m095	Δlv	Ratio, Consumer Installment Credit To Personal Income (Pct.) (TCB)	Money and Credit Quantity Aggregates
fspcom	Δln	S&P's Common Stock Price Index: Composite (1941-43=10)	Stock Prices
fspin	Δln	S&P's Common Stock Price Index: Industrials (1941-43=10)	Stock Prices
fsdxp	Δlv	S&P's Composite Common Stock: Dividend Yield (% Per Annum)	Stock Prices
fspxe	Δln	S&P's Composite Common Stock: Price-Earnings Ratio (%,Nsa)	Stock Prices
fyff	Δlv	Interest Rate: Federal Funds (Effective) (% Per Annum,Nsa)	Interest Rates and Spreads
cp90	Δlv	Cmmercial Paper Rate (AC)	Interest Rates and Spreads
fygm3	Δlv	Interest Rate: U.S.Treasury Bills,Sec Mkt,3-Mo.(% Per Ann,Nsa)	Interest Rates and Spreads
fygm6	Δlv	Interest Rate: U.S.Treasury Bills,Sec Mkt,6-Mo.(% Per Ann,Nsa)	Interest Rates and Spreads
fygt1	Δlv	Interest Rate: U.S.Treasury Const Maturities,1-Yr.(% Per Ann,Nsa)	Interest Rates and Spreads
fygt5	Δlv	Interest Rate: U.S.Treasury Const Maturities,5-Yr.(% Per Ann,Nsa)	Interest Rates and Spreads
fygt10	Δlv	Interest Rate: U.S.Treasury Const Maturities,10-Yr.(% Per Ann,Nsa)	Interest Rates and Spreads
fyaaac	Δlv	Bond Yield: Moody's Aaa Corporate (% Per Annum)	Interest Rates and Spreads
fybaac	Δlv	Bond Yield: Moody's Baa Corporate (% Per Annum)	Interest Rates and Spreads
scp90	lv	cp90-fyff (AC)	Interest Rates and Spreads
sfygm3	lv	fygm3-fyff (AC)	Interest Rates and Spreads
sfygm6	lv	fygm6-fyff (AC)	Interest Rates and Spreads
sfygt1	lv	fygt1-fyff (AC)	Interest Rates and Spreads
sfygt5	lv	fygt5-fyff (AC)	Interest Rates and Spreads
sfygt10	lv	fygt10-fyff (AC)	Interest Rates and Spreads
sfyaaac	lv	fyaaac-fyff (AC)	Interest Rates and Spreads
sfybaac	lv	fybaac-fyff (AC)	Interest Rates and Spreads
exrus	Δln	United States;Effective Exchange Rate(Merm)(Index No.)	Exchange Rates
exrsw	Δln	Foreign Exchange Rate: Switzerland (Swiss Franc Per U.S.$)	Exchange Rates
exrjan	Δln	Foreign Exchange Rate: Japan (Yen Per U.S.$)	Exchange Rates
exruk	Δln	Foreign Exchange Rate: United Kingdom (Cents Per Pound)	Exchange Rates
exrcan	Δln	Foreign Exchange Rate: Canada (Canadian $ Per U.S.$)	Exchange Rates
pwfsa	Δ^2ln	Producer Price Index: Finished Goods (82=100,Sa)	Price Indexes
pwfcsa	Δ^2ln	Producer Price Index: Finished Consumer Goods (82=100,Sa)	Price Indexes
pwimsa	Δ^2ln	Producer Price Index:I ntermed Mat.Supplies & Components(82=100,Sa)	Price Indexes
pwcmsa	Δ^2ln	Producer Price Index: Crude Materials (82=100,Sa)	Price Indexes
psccom	Δ^2ln	Spot market price index: bls & crb: all commodities(1967=100)	Price Indexes
psm99q	Δ^2ln	Index Of Sensitive Materials Prices (1990=100)(Bci-99a)	Price Indexes
pmcp	lv	Napm Commodity Prices Index (Percent)	Price Indexes
punew	Δ^2ln	Cpi-U: All Items (82-84=100,Sa)	Price Indexes

Mnemonic	Transf.	Description	Category
pu83	Δ^2ln	Cpi-U: Apparel & Upkeep (82-84=100,Sa)	Price Indexes
pu84	Δ^2ln	Cpi-U: Transportation (82-84=100,Sa)	Price Indexes
pu85	Δ^2ln	Cpi-U: Medical Care (82-84=100,Sa)	Price Indexes
puc	Δ^2ln	Cpi-U: Commodities (82-84=100,Sa)	Price Indexes
pucd	Δ^2ln	Cpi-U: Durables (82-84=100,Sa)	Price Indexes
pus	Δ^2ln	Cpi-U: Services (82-84=100,Sa)	Price Indexes
puxf	Δ^2ln	Cpi-U: All Items Less Food (82-84=100,Sa)	Price Indexes
puxhs	Δ^2ln	Cpi-U: All Items Less Shelter (82-84=100,Sa)	Price Indexes
puxm	Δ^2ln	Cpi-U: All Items Less Medical Care (82-84=100,Sa)	Price Indexes
gmdc	Δ^2ln	Pce, Impl Pr Defl:Pce (1987=100)	Price Indexes
gmdcd	Δ^2ln	Pce, Impl Pr Defl:Pce; Durables (1987=100)	Price Indexes
gmdcn	Δ^2ln	Pce, Impl Pr Defl:Pce; Nondurables (1996=100)	Price Indexes
gmdcs	Δ^2ln	Pce, Impl Pr Defl:Pce; Services (1987=100)	Price Indexes
ces275	Δ^2ln	Avg Hourly Earnings of Prod or Nonsup Workers On Private Nonfarm Payrolls - Goods-Producing	Average Hourly Earnings
ces277	Δ^2ln	Avg Hourly Earnings of Prod or Nonsup Workers On Private Nonfarm Payrolls - Construction	Average Hourly Earnings
ces278	Δ^2ln	Avg Hourly Earnings of Prod or Nonsup Workers On Private Nonfarm Payrolls - Manufacturing	Average Hourly Earnings
hhsntn	Δlv	U. Of Mich. Index Of Consumer Expectations(Bcd-83)	Miscellaneous

List of the variables enter into the Composite Leading Index of the Conference Board.

Mnemonic	Transf.	Description	Category
a0m001	lv	Average weekly hours, manufacturing	Employment and Hours
a0m005	Δln	Average weekly initial claims for unemployment insurance	Employment and Hours
a0m008	Δln	Manufacturers' new orders, consumer goods and materials	Orders
a0m027	Δln	Vendor performance, slower deliveries diffusion index	Orders
pmdel	lv	Manufacturers' new orders, nondefense capital goods	Orders
hsbr	ln	Building permits, new private housing units	Housing Starts and Sales
fspcom	Δln	Stock prices, 500 common stocks	Stock Prices
fm2dq	Δln	Money supply, M2	Money and Credit Quantity Aggregates
sfygt10	lv	Interest rate spread, 10-year Treasury bonds less Federal funds (%)	Interest Rates and Spreads
hhsntn	Δlv	Index of consumer expectations	Miscellaneous

Appendix B

Variables selection results

Table 4. Variables selection statistics, industrial production index growth rates forecasting.

	1m	3m	6m	12m	1m	3m	6m	12m
	Hard thresholding (1.28)				Hard thresholding (1.65)			
Average number of selected variables	60,85	60,78	57,81	55,64	47,47	47,38	46,86	44,80
Number of variables selected with frequency 80% and more	23	24	26	26	18	20	22	17
Number of variables selected with frequency 20% and less	29	29	43	40	44	47	55	54
10 most frequently selected variables	pmno* pmp* sfygm6* sfygm3* sfygt1 pmi sfygt5 sfygt10 a0m005 hsfr	pmno* sfygm6* sfygm3* a0m005 sfygt1 pmp sfygt5 sfygt10 lhel lhelx	a0m005* pmno* fspin* sfygm6 fsdxp fspcom hhsntn sfygm3 sfygt5 pmp	sfyaaac sfygt10 sfygt5 sfybaac fsdxp sfygm3 sfygm6 lhelx fspin pmno	pmp* pmno* sfygt1 sfygm3 sfygm6 pmi hsfr sfygt5 a0m005 ces003	pmno* sfygm3 sfygm6 sfygt1 sfygt10 sfygt5 pmp a0m005 lhelx fm2dq	pmno* sfygm6 hhsntn sfygt10 sfygm3 sfygt5 fspin sfyaaac fsdxp lhelx	sfyaaac sfygt10 sfybaac sfygt5 sfygm6 sfygm3 lhelx fm2dq fsdxp fspin
	Hard thresholding (2.58)				Soft thresholding			
Average number of selected variables	22,17	25,04	26,78	25,53	4,46	10,13	10,02	10,88
Number of variables selected with frequency 80% and more	2	5	10	7	1	1	0	0
Number of variables selected with frequency 20% and less	88	86	83	85	122	109	109	109
10 most frequently selected variables	pmno pmp sfygm6 pmi sfygm3 lhelx sfygt1 hsbr sfygt5 lhel	pmno fm2dq sfygm6 sfyaaac sfygm3 pmp sfybaac sfygt10 sfygt1 hsbr	lhelx sfyaaac pmno sfygt10 sfygm6 sfygt5 sfygm3 sfybaac sfygt1 fm2dq	sfygt10 sfyaaac sfybaac sfygt5 sfygm3 sfygm6 fm2dq lhel fsdxp fspcom	pmno lhel ces033 a0m005 pmp sfygm3 hsmw hsne lhelx ips13	pmno sfygm3 hsbr ces033 lhel hhsntn a0m005 lhelx fclbmc hsmw	pmno hsbr fm2dq sfygm3 lhel ces033 pmcp sfygm6 hsmw fclbmc	fclbmc fm2dq pmcp sfybaac sfyaaac hsbr pmno sfygm6 sfygt10 sfygm3

Variables selected with frequency 100% are marked with asterisk (*)

Table 5. Variables selection statistics, personal income less transfers growth rates forecasting.

	1m	3m	6m	12m	1m	3m	6m	12m
	Hard thresholding (1.28)				Hard thresholding (1.65)			
Average number of selected variables	55,97	66,08	64,71	60,16	43,12	54,71	54,53	50,94
Number of variables selected with frequency 80% and more	18	43	38	26	10	27	31	16
Number of variables selected with frequency 20% and less	38	33	35	42	55	45	50	49
10 most frequently selected variables	pmno*	pmno*	pmp*	lhel	pmno	pmno*	pmp	lhelx
	pmp*	pmp*	pmno	lhelx	ces015	pmp*	pmno	pmi
	ces002	ces015*	ces015	pmi	ces002	ces015	ces015	sfygm6
	ces003	ces017*	ips10	ips11	pmp	ces003	ces003	sfyaaac
	ces015	ips10	ips43	sfygm6	ces003	ces017	hsfr	sfygt5
	ces017	ces002	ces002	sfygm3	ces017	ces002	ces002	sfygm3
	ces048	ips43	ces003	sfygt10	ces048	ips10	ces017	lhel
	pmi	ces003	ips11	sfyaaac	hsfr	ips43	ips10	sfygt10
	hsfr	hsfr	ces017	sfygt5	hsbr	hsfr	ips43	sfygt1
	ips34	ips11	hsfr	sfygt1	ces033	lhur	hsbr	sfybaac
	Hard thresholding (2.58)				Soft thresholding			
Average number of selected variables	17,51	32,35	35,30	32,32	2,74	7,16	9,74	11,15
Number of variables selected with frequency 80% and more	1	6	10	7	0	0	0	0
Number of variables selected with frequency 20% and less	93	68	69	66	125	115	108	106
10 most frequently selected variables	pmno	pmno	pmno	sfyaaac	pmno	pmno	pmno	pmno
	pmp	ces015	pmp	sfygm6	ips13	hssou	hssou	sfybaac
	ces002	pmp	hsfr	pmi	ips11	hsbr	ces088	pmp
	ces015	ces002	ces015	sfygt10	sfygt1	lhelx	hsbr	ces033
	pmi	ces017	hsbr	sfygt5	a0m052	sfygt1	lhel	ces088
	ces003	hsfr	sfyaaac	sfybaac	hhsntn	ces033	ces003	pmdel
	ips13	hsbr	sfygt10	sfygt1	pmemp	lhel	pmcp	hssou
	hsbr	ces003	sfygt5	fm2dq	hsbr	hsmw	hsmw	fm2dq
	ces017	ces033	lhel	sfygm3	ips25	ces003	sfygt1	ces151
	sfygm3	lhel	sfybaac	lhel	ips299	pmcp	fyff	sfyaaac

Variables selected with frequency 100% are marked with asterisk (*)

Table 6. Variables selection statistics, nonagricultural employment growth rates forecasting.

	1m	3m	6m	12m	1m	3m	6m	12m
	Hard thresholding (1.28)				Hard thresholding (1.65)			
Average number of selected variables	56,50	61,75	55,74	55,78	41,01	49,12	44,61	44,04
Number of variables selected with frequency 80% and more	18	32	26	28	7	19	21	21
Number of variables selected with frequency 20% and less	31	32	41	41	51	50	55	58
10 most frequently selected variables	pmno*	pmno*	sfyaaac	fspcom	pmno	pmno*	sfyaaac	sfyaaac
	pmi	lhelx*	sfygt10	lhel	lhelx	lhelx	sfygt10	sfygt10
	sfygm6	sfyaaac	pmno	sfybaac	hsfr	lhel	sfygt5	sfygt5
	sfygt10	sfygt10	sfygt5	fsdxp	hssou	sfygt10	sfybaac	sfybaac
	sfyaaac	lhel	a0m005	sfyaaac	hsbr	sfyaaac	lhelx	fsdxp
	hsbr	sfygt5	lhel	sfygt10	pmi	sfybaac	pmno	pmno
	sfygm3	sfybaac	sfybaac	sfygt5	sfygm3	sfygt5	fm2dq	lhel
	sfygt5	hsfr	lhelx	lhelx	sfyaaac	hsbr	lhel	lhelx
	sfybaac	pmp	fspin	fspin	sfygm6	hsfr	fsdxp	sfygm3
	hsfr	a0m005	fspcom	pmno	ips10	a0m005	fspcom	fspcom
	Hard thresholding (2.58)				Soft thresholding			
Average number of selected variables	14,93	23,15	23,52	25,17	6,88	10,47	13,11	12,89
Number of variables selected with frequency 80% and more	0	6	8	7	0	0	0	0
Number of variables selected with frequency 20% and less	96	87	92	89	117	106	104	98
10 most frequently selected variables	pmno	sfyaaac	Sfyaaac	sfyaaac	hsbr	hsbr	lhel	sfygm6
	lhel	lhelx	Sfybaac	sfygt10	pmi	lhel	hsbr	lhelx
	lhelx	sfybaac	sfygt10	sfybaac	pmno	lhelx	pmno	hsbr
	pmp	pmno	sfygt5	sfygt5	lhel	pmi	ces033	aom001
	sfygm3	hsbr	lhelx	sfygm6	ces033	pmno	ces088	sfyaaac
	hsbr	sfygt10	fm2dq	sfygm3	aom001	aom001	sfygm6	pmno
	hsfr	lhel	pmno	pmno	pmemp	ces033	pmi	hssou
	sfygm6	hsfr	lhel	fclbmc	hssou	hssou	aom001	fm2dq
	pmi	sfygt5	pmp	pmp	hsfr	ces088	hssou	pmp
	a0m005	pmp	hsfr	lhel	hswst	sfygm3	sfyaaac	pmcp

Variables selected with frequency 100% are marked with asterisk (*)

Table 7. Variables selection statistics, manufacturing and trade sales growth rates forecasting

	1m	3m	6m	12m	1m	3m	6m	12m
	Hard thresholding (1.28)				Hard thresholding (1.65)			
Average number of selected variables	54,87	56,55	53,95	54,39	39,74	43,71	43,16	43,07
Number of variables selected with frequency 80% and more	21	24	21	20	13	15	14	13
Number of variables selected with frequency 20% and less	32	39	43	40	58	51	52	54
10 most frequently selected variables	lhel* pmno sfygt1 sfygm6 lhelx pmi ces003 sfygt5 sfygt10 sfygm3	lhelx sfygt10 sfygm6 pmno sfygt5 sfyaaac sfybaac sfygt1 hsbr fm2dq	sfygt10* sfyaaac pmno sfygm6 sfygt5 sfybaac lhelx sfygt1 pmp hssou	sfygt10* sfyaaac* sfygt5* sfybaac* sfygm6 pmno sfygt1 sfygm3 hhsntn lhel	lhel pmno sfygm6 lhelx a0m051 a0m052 sfygt1 sfygm3 sfygt10 sfyaaac	lhelx sfyaaac sfygt5 sfygt10 sfybaac pmno sfygm6 sfygt1 fm2dq hhsntn	sfyaaac sfygm6 sfybaac sfygt10 pmno sfygt5 lhelx sfygt1 pmp sfygm3	sfygt10* sfyaaac* sfybaac* sfygt5 sfygm6 sfygt1 sfygm3 pmno hhsntn scp90
	Hard thresholding (2.58)				Soft thresholding			
Average number of selected variables	16,69	21,43	23,85	25,04	1,81	4,29	6,02	6,39
Number of variables selected with frequency 80% and more	0	3	6	8	0	0	0	0
Number of variables selected with frequency 20% and less	92	86	86	80	123	119	115	115
10 most frequently selected variables	lhel pmno sfygm6 sfygm3 sfygt10 sfyaaac a0m051 sfygt5 pmp lhelx	sfyaaac sfybaac sfygt10 sfygt5 sfygm6 pmno fm2dq lhelx sfygm3 fybaac	sfyaaac sfygt10 sfygt5 sfybaac pmno sfygm6 sfygt1 fm2dq sfygm3 hsbr	sfyaaac* sfygt10 sfybaac sfygt5 sfygt1 sfygm6 sfygm3 scp90 fm2dq fclbmc	lhel a0m077 a0m051 lhu26 ips38 sfygt10 lhelx pmno sfygm3 ips299	sfyaaac sfygt10 hsbr fm2dq fybaac sfygm6 fclbmc ces033 pmno hhsntn	fm2dq fybaac sfygm6 sfyaaac sfygt10 sfybaac fclbmc pmcp hsbr hssou	fclbmc sfybaac fm2dq pmcp sfyaaac sfygt10 ces088 hsbr a0m051 hsne

Variables selected with frequency 100% are marked with asterisk (*)

Table 8. Mean squared prediction errors relative to variance, industrial production index growth rates forecasting.

				Principal component regression					Principal covariate regression				
Sample	Variance	AR	CLI	no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)	no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)
1 month													
1970-2003	74,145	0,880	**0,792**	0,754	**0,740**	0,744	0,753	0,752	0,768	0,723	0,762	**0,719**	0,742
1970-1983	128,843	0,805	**0,739**	0,680	**0,666**	0,677	0,686	0,679	0,706	0,637	0,696	**0,603**	0,647
1984-1993	33,253	1,140	**0,922**	**1,004**	**1,004**	1,005	1,023	1,040	1,053	1,023	**0,935**	1,034	1,130
1994-2003	37,694	1,021	**0,942**	0,901	0,874	**0,843**	0,848	0,855	**0,822**	0,881	0,932	1,004	0,861
3 months													
1970-2003	43,451	0,903	**0,720**	0,752	0,787	0,712	0,686	**0,676**	0,735	0,692	0,660	0,691	**0,649**
1970-1983	81,004	0,867	**0,673**	0,704	0,746	0,662	0,623	**0,613**	0,710	0,628	0,608	0,640	**0,568**
1984-1993	14,635	1,097	**0,850**	1,078	1,110	1,040	1,082	**1,014**	1,114	1,101	**1,028**	1,113	1,161
1994-2003	18,653	0,993	**0,920**	0,806	0,799	**0,773**	0,782	0,812	**0,608**	0,775	0,699	0,686	0,754
6 months													
1970-2003	31,112	1,036	**0,788**	0,854	0,714	0,683	**0,627**	**0,627**	0,829	0,713	0,687	0,662	**0,632**
1970-1983	57,356	1,004	**0,711**	0,816	0,580	0,575	0,495	**0,492**	0,740	0,584	0,589	0,565	**0,488**
1984-1993	9,691	1,118	**1,100**	**1,102**	1,412	1,243	1,321	1,309	1,516	1,581	1,466	**1,401**	1,597
1994-2003	14,675	1,179	**1,031**	0,921	1,020	0,935	**0,919**	0,949	0,880	0,866	0,723	**0,721**	0,800
12 months													
1970-2003	20,406	1,086	**0,766**	1,035	0,630	0,700	0,582	**0,565**	0,733	0,599	0,534	**0,513**	0,542
1970-1983	36,015	1,095	**0,678**	1,047	0,397	0,574	0,393	**0,358**	0,578	0,341	0,372	0,325	**0,318**
1984-1993	5,889	**0,910**	0,951	**0,905**	1,576	1,099	1,244	1,308	1,554	1,883	**1,274**	1,375	1,410
1994-2003	12,055	1,156	**1,086**	**1,068**	1,211	1,075	1,114	1,133	1,021	1,109	**0,895**	0,929	1,121

Table 9. Mean squared prediction errors relative to variance, personal income less transfers growth rates forecasting.

				Principal component regression					Principal covariate regression				
Sample	Variance	AR	CLI	no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)	no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)
1 month													
1970-2003	26,425	0,998	**0,982**	0,878	0,896	**0,867**	0,871	0,892	0,891	0,902	0,882	**0,865**	0,894
1970-1983	35,413	0,908	**0,887**	0,838	0,835	**0,827**	0,834	0,852	0,868	0,846	0,844	**0,817**	0,846
1984-1993	23,218	1,131	**1,121**	**0,948**	0,978	0,957	0,963	0,975	**0,931**	0,978	0,973	0,971	0,979
1994-2003	16,701	1,097	**1,088**	0,914	0,981	0,875	**0,867**	0,910	0,915	0,979	0,884	**0,877**	0,935
3 months													
1970-2003	13,427	1,011	**0,942**	0,801	0,875	0,776	0,769	**0,766**	0,809	0,846	0,802	0,797	**0,759**
1970-1983	19,622	0,989	**0,887**	0,821	0,926	0,783	0,776	**0,772**	0,838	0,847	0,819	0,837	**0,760**
1984-1993	7,987	1,100	**1,081**	0,742	**0,719**	0,745	0,734	0,721	0,781	**0,722**	0,773	0,742	0,738
1994-2003	9,912	1,023	**1,003**	0,808	0,877	0,797	**0,791**	0,801	0,768	0,964	0,791	**0,743**	0,790
6 months													
1970-2003	9,045	1,167	**0,992**	0,931	1,032	0,872	0,850	**0,845**	0,939	0,946	0,760	0,770	**0,836**
1970-1983	12,710	1,225	**0,920**	1,008	1,166	0,906	0,881	**0,863**	0,914	0,895	0,728	**0,711**	0,849
1984-1993	5,452	1,155	**1,136**	**0,815**	0,826	0,844	0,828	0,823	0,901	0,968	0,800	0,827	**0,774**
1994-2003	7,294	**1,051**	1,084	0,842	0,869	0,820	**0,804**	0,829	1,049	1,074	**0,823**	0,890	0,867
12 months													
1970-2003	6,228	1,191	**0,964**	1,052	0,911	0,941	0,898	**0,862**	0,969	0,908	0,888	**0,886**	0,894
1970-1983	8,658	1,158	**0,814**	1,135	0,791	0,972	0,896	**0,815**	0,952	**0,753**	0,891	0,823	0,832
1984-1993	3,105	1,187	**1,157**	0,909	1,008	**0,905**	0,941	1,016	1,034	1,045	**0,921**	1,068	1,141
1994-2003	5,732	1,311	**1,229**	0,979	1,163	0,921	**0,906**	0,908	1,001	1,217	**0,892**	0,955	0,922

Table 10. Mean squared prediction errors relative to variance, nonagricultural employment growth rates forecasting.

				Principal component regression					Principal covariate regression				
Sample	Variance	AR	CLI	no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)	no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)
1 month													
1970-2003	9,559	0,595	**0,541**	**0,505**	0,535	0,525	0,506	0,525	0,511	0,525	0,521	0,540	**0,508**
1970-1983	17,226	0,668	**0,590**	0,534	0,538	0,547	**0,520**	0,560	0,541	**0,522**	0,536	0,549	0,532
1984-1993	3,994	**0,410**	0,425	0,512	0,634	0,568	0,589	**0,477**	0,535	0,621	0,565	0,586	**0,481**
1994-2003	4,231	**0,365**	0,387	**0,346**	0,443	0,368	0,360	0,386	**0,328**	0,465	0,407	0,459	0,410
3 months													
1970-2003	7,026	0,546	**0,397**	0,417	0,424	0,419	**0,401**	0,423	0,362	0,401	**0,356**	0,386	0,373
1970-1983	12,309	0,647	**0,450**	0,464	**0,425**	0,454	**0,425**	0,452	0,381	0,403	**0,371**	0,411	0,377
1984-1993	3,020	0,300	**0,275**	**0,343**	0,493	0,412	0,423	0,385	0,371	0,487	**0,355**	0,366	0,375
1994-2003	3,428	0,263	**0,251**	**0,255**	0,370	0,259	0,270	0,324	**0,266**	0,327	0,290	0,287	0,364
6 months													
1970-2003	6,082	0,658	**0,510**	0,539	**0,458**	0,499	0,495	0,496	0,468	0,485	0,441	**0,415**	0,421
1970-1983	10,317	0,790	**0,589**	0,608	**0,444**	0,534	0,529	0,535	0,493	0,450	0,444	**0,407**	0,419
1984-1993	2,723	0,375	**0,347**	**0,439**	0,559	0,479	0,481	0,485	**0,489**	0,683	0,525	0,504	0,471
1994-2003	3,220	0,315	**0,305**	**0,331**	0,460	0,376	0,370	0,345	**0,353**	0,498	0,376	0,389	0,409
12 months													
1970-2003	4,773	0,811	**0,685**	0,712	0,628	0,617	0,540	**0,501**	0,555	0,612	0,480	0,472	**0,468**
1970-1983	7,523	0,984	**0,790**	0,817	0,582	0,587	0,502	**0,460**	0,470	0,426	**0,357**	0,390	0,371
1984-1993	2,380	**0,550**	0,590	**0,566**	0,848	0,825	0,833	0,733	1,012	1,174	0,905	0,819	**0,808**
1994-2003	2,911	0,424	**0,405**	0,486	0,663	0,599	**0,474**	0,498	**0,527**	0,901	0,629	0,529	0,592

Table 11. Mean squared prediction errors relative to variance, manufacturing and trade sales growth rates forecasting

Sample	Variance	AR	CLI	Principal component regression					Principal covariate regression				
				no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)	no selection	soft	hard (1.28)	hard (1.65)	hard (2.58)
1 month													
1970-2003	154,498	0,997	**0,961**	0,936	0,992	**0,892**	0,899	0,931	**0,892**	1,017	0,914	0,929	0,960
1970-1983	198,303	1,027	**0,974**	0,954	1,000	0,870	**0,859**	0,909	0,914	1,033	**0,852**	0,854	0,899
1984-1993	137,014	1,025	**0,952**	0,939	0,973	**0,869**	0,880	0,915	**0,832**	0,979	0,875	0,964	0,911
1994-2003	110,061	**0,889**	0,940	**0,889**	0,997	0,978	1,023	1,009	**0,912**	1,028	1,122	1,079	1,181
3 months													
1970-2003	46,697	1,037	**0,908**	0,962	0,919	0,889	**0,884**	0,886	0,880	0,916	0,857	0,850	**0,834**
1970-1983	82,512	1,034	**0,868**	0,924	0,812	0,824	0,804	**0,788**	0,809	0,785	0,782	0,737	**0,718**
1984-1993	22,024	1,042	**0,969**	**1,077**	1,266	**1,077**	1,080	1,113	**1,024**	1,393	1,093	1,172	1,113
1994-2003	20,480	**1,056**	1,082	1,064	1,156	**1,062**	1,134	1,209	1,135	1,149	**1,036**	1,155	1,200
6 months													
1970-2003	27,984	1,109	**0,945**	1,047	**0,759**	0,874	0,868	0,778	0,805	0,806	**0,755**	0,781	0,803
1970-1983	52,342	1,127	**0,890**	1,034	**0,608**	0,807	0,780	0,642	0,662	0,630	**0,595**	0,607	0,627
1984-1993	10,634	**0,994**	1,202	1,085	1,286	**1,056**	1,073	1,164	1,335	1,379	**1,263**	1,377	1,303
1994-2003	10,300	1,102	**1,086**	**1,106**	1,316	1,183	1,310	1,376	**1,306**	1,507	1,406	1,445	1,587
12 months													
1970-2003	16,707	1,124	**0,877**	1,078	**0,578**	0,783	0,787	0,609	0,688	0,548	0,628	0,574	**0,509**
1970-1983	32,188	1,148	**0,820**	1,093	**0,439**	0,715	0,707	0,476	0,551	0,411	0,476	0,412	**0,396**
1984-1993	5,468	**0,896**	1,143	0,890	0,891	0,875	**0,775**	0,935	1,370	1,079	1,034	1,071	**0,949**
1994-2003	5,111	1,168	**1,124**	**1,156**	1,568	1,341	1,580	1,520	1,220	1,264	1,634	1,569	**1,095**