Dynamic Factor Models and Forecasting Finnish Macroeconomic Variables

Paolo Fornaro

University of Helsinki

Faculty Of Social Sciences

Economics

Master's Thesis

May 2011

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto – Fakultet/Sektion – Faculty	Laitos – Institution – Department
Social Science	Economics

Tekijä – Författare – Author
Paolo Fornaro

Työn nimi – Arbetets titel – Title
Dynamic Factor Models and Forecasting Finnish Macroeconomic Variables

Oppiaine – Läroämne – Subject
Economics:Econometrics

Työn laji – Arbetets art – Level	Aika – Datum – Month and year	Sivumäärä – Sidoantal – Number of pages
Master's Thesis	May 2011	63

Tiivistelmä – Referat – Abstract

In recent years, thanks to developments in information technology, large-dimensional datasets have been increasingly available. Researchers now have access to thousands of economic series and the information contained in them can be used to create accurate forecasts and to test economic theories. To exploit this large amount of information, researchers and policymakers need an appropriate econometric model. Usual time series models, vector autoregression for example, cannot incorporate more than a few variables. There are two ways to solve this problem: use variable selection procedures or gather the information contained in the series to create an index model. This thesis focuses on one of the most widespread index model, the dynamic factor model (the theory behind this model, based on previous literature, is the core of the first part of this study), and its use in forecasting Finnish macroeconomic indicators (which is the focus of the second part of the thesis). In particular I forecast economic activity indicators (e.g. GDP) and price indicators (e.g. consumer price index), from 3 large Finnish datasets. The first dataset contains a large series of aggregated data obtained from the Statistics Finland database. The second dataset is composed by economic indicators from Bank of Finland. The last dataset is formed by disaggregated data from Statistic Finland, which I call micro dataset. The forecasts are computed following a two steps procedure: in the first step I estimate a set of common factors from the original dataset. The second step consists in formulating forecasting equations including the factors extracted previously. The predictions are evaluated using relative mean squared forecast error, where the benchmark model is a univariate autoregressive model. The results are dataset-dependent. The forecasts based on factor models are very accurate for the first dataset (the Statistics Finland one), while they are considerably worse for the Bank of Finland dataset. The forecasts derived from the micro dataset are still good, but less accurate than the ones obtained in the first case. This work leads to multiple research developments. The results here obtained can be replicated for longer datasets. The non-aggregated data can be represented in an even more disaggregated form (firm level). Finally, the use of the micro data, one of the major contributions of this thesis, can be useful in the imputation of missing values and the creation of flash estimates of macroeconomic indicator (nowcasting).

Avainsanat – Nyckelord – Keywords
Forecasting, Factor Model, Large Datasets, Micro Data

Contents

1 Introduction

In recent years, thanks to developments in information technology, large-dimensional datasets have become increasingly available. Researchers now have access to thousands of economic series and the information contained in them can be used to create accurate forecasts, as well as to test economic theories. To benefit from this large amount of information, researchers and policymakers need an appropriate econometric model. Usual time series models, vector autoregression for example, cannot incorporate more than a few variables. If the number of parameters to estimate is large with respect to the number of observations, the model would run into a scarce degrees of freedom problem, typical of regression-based analysis. This consists in the difficulty for the model, to estimate the various parameters, which would be characterized by large variances. Furthermore, models containing more variates than observations cannot be estimated in traditional frameworks. These limitations do not allow for the full use of information included in large datasets.

There are two ways to solve the degrees of freedom problem: use variable selection procedures, or create an index model from the information contained in the series. The first methodology involves selecting, from a large set of series, the most relevant variables for the econometric model. Examples of variables selection procedures are the general-to-specific variable selection, discussed by Hendry (1995), the simulated annealing (Kapetanios, 2007), genetic algorithms(Dorsey and Mayer, 1995), boosting (Buhlmann, 2006) and the least absolute shrinkage and selection operator by Robert Tibshirani (1996). The problem with these procedures is that the model is still based only on the few chosen variables, and much of the information carried by the large dataset would be lost.

The second methodology entails in using all the information available in the dataset to create a handful of predictors. This principle is implemented in two main classes of statistical models. The first is the principal component analysis (PCA) and the second is the factor analysis (FA).

In this thesis I focus on dynamic factor models and on their ability to forecast Finnish macroeconomic variables. While forecasts obtained by models based on common factors have been widely studied in the last ten years, a forecasting

experiment with large Finnish datasets has not been done. In the theoretical part of this paper I present the factor model, its dynamic version and various developments. In particular, I introduce the model by Stock and Watson (2002b) and by Forni, Hallin, Lippi and Reichlin (2000). After the theoretical discussion I present a small survey on the various economic applications in which dynamic factor models have been used. Finally, a forecasting analysis for the Finnish economy is conducted. In this analysis, I estimate a set of factors from three different datasets.

The first dataset includes a large series of macroeconomic variables gathered from different Statistics Finland databases (StatFin). The second dataset consists of a set of macroeconomic indicators obtained from the Bank of Finland (BOF) databases, which spans a longer time period compared to the other two datasets used in this thesis. The final dataset contains a large number of micro variables: by micro I mean that the variables are in a less aggregated form than the other two datasets. The use of a dataset containing such disaggregated variables is the other unique element of this research, and the aim is to shed light on the ability of the factors extracted from micro data to produce reliable forecasts. I include more detailed descriptions of the datasets in separate sections.

In the empirical analysis, the static principal component method by Stock and Watson (2002b) is used as an estimation method of the factors. Factors are also extracted using the method formulated by Forni, Hallin, Lippi and Reichlin (2000). The resulting factors are very closely correlated with the common components estimated by means of the Stock and Watson (SW) method. This method is easier to implement, hence I compute the forecasts using this model. The estimated factors are used to compute forecasts for selected series, which are then compared to forecasts obtained through the use of a number of benchmark models (namely an autoregressive model and a vector autoregressive model). The results are data-dependent. For the StatFin and the micro dataset the factor models perform well in terms of forecasting, but they fail to create accurate predictions for the BOF dataset. A very interesting result related to the StatFin forecasts is the ability of the factor models to create accurate prediction for the price indicators. This finding is in sharp contrast with the past literature. The ability of the

factors extracted from the micro data to create relatively good forecasts is another peculiar result of this study.

The thesis is structured as follows: section 2 introduces the static basic factor model and the approximate factor model, section 3 focuses on the dynamic factor model and section 4 describes two of the most important "second generation" factor models, the SW(2002b) and the Forni et al.(2000) models. I also include a short summary of the model by Kapetanios and Marcellino (2006). In section 5 the main economic applications that use dynamic factor models are listed, sections 6, 7 and 8 include the forecasting experiments for the three datasets considered in this work, and, finally, in section 9 I draw some conclusions and suggest possible developments which could be derived from this work.

Part I

Theoretical Background

In this part I present an overview of the factor model and its developments. The first model under consideration is the basic, static factor model. After that I introduce the dynamic factor model and the theoretical background is concluded by introducing two of the most important extensions of the dynamic factor model, namely the model by SW (2002b) and the model by Forni, Hallin, Lippi and Reichlin (2000).

2 The Basic Model

This section follows the presentation by D.N. Lawley and A.E. Maxwell(1962). The main assumption of factor analysis (in this case the r-factor model) is that for a vector of variables $y_i = \left[\; y_1, ..., y_N \; \right]'$, the following representation holds.

$$y_i = \sum_{k=1}^{r} \lambda_{ik} f_k + e_i \tag{2.1}$$

(i=1,2,...,N)

where f_k is the k-th common factor, r is specified, and u_i is the idiosyncratic component of variable y_i. The error terms e_i are assumed to be mutually uncorrelated, with $E(e_i) = 0$ and $E(e_i, e_i) = \sigma_i^2$. It is also assumed that $E(f_k) = 0$ and, without loss of generality, $E(f_k, f_k) = 1$. A further assumption is that $E(f_k, e_i) = 0$ for all i and k. Here λ_{ik} and σ_i^2 have to be estimated. It is important to remember that also f_k is unknown, and it can be interesting to estimate it.

2.1 Factor Loadings Estimation

One of the main interest of FA is the estimation of the factor loadings in (2.1). There has been numerous methods to estimate the loadings, for example the centroid method (also called simple summation method), the ordinary least square

method, the weighted least square method and the method of maximum likelihood. Here the focus is on the maximum likelihood method.

To show this estimation method, it is useful to start from equation (2.1). It is assumed that y_i follows a multivariate normal distribution. The variance-covariance matrix associated to y is denoted $\mathbf{C}$=$[c_{ij}]$, and it is of order N. The factors are assumed to be orthogonal and uncorrelated. r, the number of factors, must not be too large, and a usual condition is that $(N + r) < (N - r)^2$. Variance of y_i is then given by:

$$c_{ii} = \sum_{k=1}^{r} \lambda_{ik}^2 + \sigma_i^2,$$
$$c_{ij} = \sum_{k=1}^{r} \lambda_{ik}\lambda_{jk}$$

Another representation for these two equations is the following:

$$\mathbf{C} = \mathbf{L}\mathbf{L}' + \mathbf{V}, \tag{2.2}$$

where $\mathbf{L}$=$[l_{ik}]$ is a N x r matrix of loadings and $\mathbf{V}$ is a diagonal matrix containing σ_i^2 on its diagonal. Let $\mathbf{A}$=$[a_{ij}]$ be the sample covariance matrix of y_i, whose elements are the sample estimates of c_{ii} and c_{ij}. The likelihood function is given by:

$$L = -\frac{1}{2} n log|\mathbf{C}| - \frac{1}{2} n \sum_{i,j} a_{ij} c^{ij} \tag{2.3}$$

So the estimated λ's are obtained by maximizing (2.3) with respect to λ_{ik} and σ_i^2. To provide an unique solution it is necessary to choose a $\mathbf{L}$ matrix, such that

$$\mathbf{J} = \mathbf{L}'\mathbf{V}^{-1}\mathbf{L}$$

is diagonal. To solve the maximization problem here presented, it is needed to equate to zero the partial derivatives of (2.3) with respect to λ_{ik} and σ_i^2. We then

get

$$\frac{\partial L}{\partial \lambda_{ik}} = -n(\sum_{j} \lambda_{jk} c^{ji} - \sum_{j,u,w} \lambda_{jk} c^{ju} a_{uw} c^{wi})$$

and

$$\frac{\partial L}{\partial \sigma_i^2} = -\frac{1}{2} n(c^{ii} - \sum_{u,w} c^{iu} a_{uw} c^{wi})$$

From these equations it is not easy to obtain a direct solution, but they can be simplified to get $\hat{c}_{ii} = a_{ii}$ and

$$\sigma_i^2 = a_{ii} - \sum_{k=1}^{r} \hat{\lambda}_{ik}^2 \tag{2.4}$$

(i=1,...,N)

Another equation obtained is

$$\hat{\mathbf{L}}' = \hat{\mathbf{J}}^{-1} \hat{\mathbf{L}}' \hat{\mathbf{V}}^{-1} (\mathbf{A} - \hat{\mathbf{V}}) \tag{2.5}$$

Equations (2.4) and (2.5) are then solved by iteration, where fairly good initial estimates of the loadings would render the iteration procedure faster.

2.2 Factor Scores Estimation

As mentioned before, it might be crucial to have an estimate of the unobservable factors. For example, in the empirical part of this thesis the focus is on the estimation of the factors underlying the various datasets. These factors are then used to compute forecasts for various economic indicators. This kind of application takes only into account the estimated factors themselves and the factor loadings are not necessary for the analysis.The following factor scores estimation procedure is due to Bartlett(1938).

The main idea is to minimize $\sum_i e^2/\sigma_i^2$, the sum of squared standardized residuals.

The previous sum can be rewritten as

$$\sum_{i=1}^{N}(y_i - \sum \lambda_{ik} f_k)^2/\sigma_i^2,$$

which must minimized with respect to $f_1, \cdots, f_r$. The minimization of the previous equation leads to

$$\sum_{i,s}(\lambda_{ik}\lambda_{is}/\sigma_i^2)f_s = \sum_{i}(\lambda_{ik}y_i/\sigma_i^2)$$

for $k = 1, \cdots, r$, where the estimate of f_k are denoted by $\hat{f}_k$. Rewriting these equations we get the final formula

$$\begin{aligned}(\mathbf{L}'\mathbf{V}^{-1}\mathbf{L})\hat{\mathbf{f}} &= \mathbf{L}'\mathbf{V}^{-1}\mathbf{y} \\ \hat{\mathbf{f}} &= \mathbf{J}^{-1}\mathbf{L}'\mathbf{V}^{-1}\mathbf{y}\end{aligned}$$

2.3 Factor Interpretation

Once the estimation of the factors and of the corresponding loadings has been completed, it is interesting to assign an interpretation to the factors. However, this is not a necessary step of FA. For example in SW (2002a), the authors use FA to create a number of indexes to improve macroeconomic forecasts. While they focus on the forecasts, they do not try to give an interpretation to the obtained factors. In certain applications, though, it is a key issue to obtain a meaningful identification of the factors. For example, if the researcher is trying to find few common factors that explain a large macroeconomic dataset, containing numerous variables, the interest may be focused on seeing if the factors obtained represent the common business cycle of the original variables. The coherence function between the common factor and the variables may then be checked, to see if there is a peak corresponding to the business cycle frequencies.

The initial estimates of the factor loadings may be difficult to interpret. A factor can have a positive factor loading on a variable, and a negative loading on others. It is also possible to have many loadings, which would increase the parameters necessary to describe the data. Because of these reasons, it is common to apply a linear transformation to the initial set of loadings. This transformation is called

factor rotation and it has three main objectives. The first one is to reduce as much as possible the number of negative loadings, which are difficult to interpret. The second one is to reduce to zero, or near zero, as many loadings as possible, in order to reduce the number of parameters in the model. Finally, factor rotation is used to separate, on different factors, loadings that contrast with each other. One problem of factor rotation is that it is based on a subjective assessment, although techniques to achieve a unique set of loadings have been developed.

It is useful to have a look at a practical example. Suppose we want to estimate a factor model for a series of school subjects. In other words, we are interested in knowing the factors, and their loading coefficients, affecting the performance in school. An initial factor loading estimates, for a two factors model, is showed in the following table.

Subjects	1st Factor	2nd Factor
Maths	0.78	-0.15
Chemistry	0.87	-0.59
Biology	0.81	-0.42
Physics	0.67	-0.33
Philosophy	0.63	0.34
History	0.77	0.27
Literature	0.82	0.43

Table 1: Loadings on Three Factors for Seven School Subjects

Looking at the table it is easy to give an interpretation to the first factor. Having a positive factor load on all subjects, this factor can be interpreted as "overall intelligence". Once the effect of the first factor on the variables is removed, we get that the second factor has positive and negative loadings. In this case, also this factor can be easily interpreted. The factor number two has positive loadings on humanistic subject, while negative loadings on scientific ones. Hence, factor two can be called "propensity to humanistic subjects against scientific subjects". In this case, negative factor loadings were easily interpreted, but this is not always the case.

2.4 Approximate Factor Model

Before introducing the dynamic factor model, it is interesting to introduce a class of models where some of the restrictive assumptions of the basic factor model are relaxed. The main element, of the approximate factor model, is that the number of variables N tends to infinity. If this assumption is true then it is possible to allow for weak serial correlation of the idiosyncratic terms. We must remark, though, that in this case the idiosyncratic terms are assumed to be generated by a stationary ARMA process, while random walks are ruled out. This class of models also accepts heteroskedastic error terms, and even weak correlation among the factors and the idiosyncratic terms. Finally, in this model, the factors contribute to the variables with a similar order of magnitude, ruling out the possibility of factors contributing to only a limited number of variables. For this class of models Bai and Ng (2002) have formulated an information criteria, for datasets across section and time, for N and T tending to infinity. We define

$$V(K) = (NT)^{-1} \sum_{t=1}^{T} \hat{u}_t' \hat{u}_t,$$

the overall sum of squared errors of a K factor model. Then the information criteria is

$$IC_{p2}(k) = log[V(k)] + k(\frac{N+T}{NT}) log[min\{N, T\}].$$

$\hat{k}$ is obtained by minimizing the information criteria in the range $k = 0, 1, ..., kmax$ where $kmax$ is a pre-specified upper bound. In this thesis I will not use this information criteria. The main reason for this decision stands in the fact that previous literature, on forecasting based on factor models, seems agnostic about the determination of factors for forecasting. The forecasting experiments are usually based on using various models including different numbers of factors. A common finding is that only few factors contribute in the creation of accurate forecasts. I follow this practice, using in the forecasting equations only three factors.

3 The Dynamic Factor Model

Papers from Geweke (1977) and Sargent and Sims (1977) introduced the seminal idea of dynamic factor models. The model I present here is the basic dynamic factor model by Sargent and Sims (1977). The main idea of this model is that the observation t of a dataset can be modeled as the sum of a number of common factors, the lags of these common components and an idiosyncratic component. This model can be summed up in the following equation

$$y_t = \Lambda_0 f_t + \Lambda_1 f_{t-1} + ... + \Lambda_m f_{t-m} + e_t \tag{3.1}$$

where $\Lambda_0, ..., \Lambda_m$ are N x r matrices and f_t is a vector of r factors. Finally e_t is the vector of idiosyncratic components, which are assumed to be independent stationary processes. This means that these components are uncorrelated to both leads and lags of the common factors and to the other idiosyncratic components. The estimation of the loading matrices, of the factors and of the rest of the parameters of the model, can be achieved by a particular maximum likelihood technique called Kalman filter. The essential features of the Kalman filter are presented in the next subsection.

3.1 Kalman Filter

I use the presentation formulated by Hamilton (1994).The Kalman filter starts from a particular dynamic model called state-space system. It consists in the following system of equations

$$s_{t+1} = F \cdot s_t + v_{t+1} \tag{3.2}$$

$$y_t = A' \cdot x_t + H' \cdot s_t + w_t \tag{3.3}$$

where s_t is a r x 1 vector, F is r x r matrix, v_t is again an r x 1. y_t is N x 1, A'

is N x k, x_t is k x 1, H' is N x r and finally w_t is N x 1. To complete the model I define $E(v_t, v'_\tau) = Q$ for $t = \tau$, while it is zero for $t \neq \tau$, and $E(w_t, w'_\tau) = R$ for $t = \tau$, while it is zero for $t \neq \tau$. It is also assumed that $E(v_t, w'_t) = 0$ Equation (3.2) is called state equation and equation(3.3) is called observation equation.

It is assumed that values for $y_1, ..., y_T$ and $x_1, ..., x_T$ are known. The Kalman filter method allows to find optimal linear projection of $\hat{s}_{t|t-1}$ and $\hat{y}_{t|t-1}$, using information contained in (x_t, Y_{t-1}) where $Y_{t-1} \equiv (y'_{t-1}, ..., y'_1, x'_{t-1}, ..., x'_1)'$. The Kalman filter algorithm is started by defining the unconditional mean and variance of s_1:

$$s_{1|0} = E(s_1)$$
$$P_{1|0} = E[(s_1 - E(s_1))(s_1 - E(s_1))']$$

where $P_{1|0}$ is called mean square error (MSE) matrix. It can be shown that the best linear forecast of $s_{t+1|t}$ is given by

$$\hat{s}_{t+1|t} = F\hat{s}_{t|t-1} + FP_{t|t-1}H(H'P_{t|t-1}H + R)^{-1}(y_t - A'x_t - H'\hat{s}_{t|t-1}) \tag{3.4}$$

with an associated MSE

$$P_{t+1|t} = F[P_{t|t-1} - P_{t|t-1}H(H'P_{t|t-1}H + R)^{-1}H'P_{t|t-1}]F' + Q \tag{3.5}$$

The Kalman filter methodology is given by iterating (3.4) and (3.5) for $t = 1, ..., T$. The forecast of y_{t+1} is given by

$$\hat{y}_{t+1|t} \equiv \hat{E}(y_{t+1}|x_{t+1}, Y_t) = A'x_{t+1} + H'\hat{s}_{t+1|t} \tag{3.6}$$

with associated MSE

$$E[(y_{t+1} - \hat{y}_{t+1|t})(y_{t+1} - \hat{y}_{t+1|t})'] = H' P_{t+1|t} H + R \tag{3.7}$$

In the process just described it is assumed the F, Q, A, H and R are known. In practice these parameters are unknown. Actually, as in the dynamic factor analysis, an exact estimation of these parameters could be the central purpose of using the Kalman filter. To obtain these parameters, we need to know the density function of $y_t|x_t, Y_{t-1}$. If s_1 and (w_t, v_t) for $t = 1, ..., T$ are Gaussian then $y_t|x_t, Y_{t-1}$ is Gaussian with mean and variance given by (3.6) and (3.7).

$$y_t|x_t, Y_{t-1} \sim N((A'x_{t+1} + H'\hat{s}_{t+1|t}), (H' P_{t+1|t} H + R)).$$

Based on these assumptions the density function for the random variable $\mathbf{Y}$ is:

$$\begin{aligned} f_{\mathbf{Y}_t|\mathbf{X}_t,T_t}(y_t|x_t, Y_{t-1}) =& (2\pi)^{-n/2} |H' P_{t|t-1} H + R|^{-1/2} \\ & \times \exp\{-\frac{1}{2}(y_t - A'x_t - H'\hat{s}_{t|t-1})' \\ & \times (H' P_{t|t-1} H + R)^{-1} \\ & \times (y_t - A'x_t - H'\hat{s}_{t|t-1})\} \end{aligned}$$

For $t = 1, ..., T$.

It is then simple to construct the sample likelihood function

$$\sum_{t=1}^{T} log f_{Y|X_t, Y_{t-1}}(y_t|x_t, Y_{t-1}) \tag{3.8}$$

and maximize it with respect the unknown parameters of F, Q, A, H and R. To obtain the density function of $\mathbf{Y}$ we need to insert some initial values for the parameters we want to estimate. These initial estimates can be derived from OLS regression of y_t on x_t and s_t, getting initial estimates of F, Q, A, H and R. We

can include these estimates in the iteration process represented by equation (3.4) to equation (3.7). From this iteration process we get the sequences $\{\hat{s}_{t|t-1}\}_{t=1}^{T}$ and $\{\hat{P}_{t|t-1}\}_{t=1}^{T}$, which then are used to calculate the density function, and the related log-likelihood function. This procedure is then iterated until (3.9) is maximized with respect to F, Q, A, H and R.

3.2 Estimation of Dynamic Factor Models

The use of Kalman filter to estimate the parameters in (3.1), requires us to rewrite the model in a state space form. One possible way to write the model is as following:

$$\begin{aligned} y_{Nt} &= \Lambda f_t + e_t \\ f_t &= A f_{t-1} + u_t \end{aligned} \tag{3.9}$$

For t=1,...,T.

Here y_{Nt} is a vector of stationary zero-mean variables, f_t is a r dimensional vector of factors at t and u_t and ϵ_t are stationary and mutually uncorrelated. For the use of Kalman filter another assumption is needed, namely that u_t and ϵ_t are Gaussian.

We can then apply the Kalman filter algorithm to (3.9), getting estimates of Λ and A. Among the benefits of the Kalman filter is the fact that we can have variables sampled at different frequencies and that the state-space representation allows to write very flexible models. The disadvantages of the Kalman filter consists in the fact that we have to assume errors following a normal distribution. Another problem is that for very large datasets, this algorithm is very computationally intensive, because it involves maximum likelihood methods. This kind of problem can be solved by so called "second generation" dynamic factor model, where the estimation techniques can handle very large datasets.

4 Developments of the Dynamic Factor Model

In the last few years a series of papers by SW (2002b), Forni, Lippi, Hallin and Reichlin (2000), Forni and Lippi (2001) had the main objective to develop new versions of the dynamic factor model, presented in the previous section. These new models try to relax some of the restrictions of the basic model, while trying to formulate new estimation methods, that are more time-efficient and that can allow the use of larger datasets. In the next subsections I will present the models from the previously cited authors. These kind of models are also addressed as "second generation" models. Finally I will describe briefly the model by Kapetanios and Marcellino (2006), which unifies the state-space representation, typical of "first generation" models, together with the ability to handle large dataset.

4.1 The Model by Stock and Watson(2002b)

The main idea of this model is to combine the approximate factor model and the dynamic factor model. The approximate factor model is an extension of the static model, which allows heteroskedasticity and weak serial and cross-correlation of the idiosyncratic terms. In the paper, the authors use the estimated factors to create out-of-sample forecasts. The model starts by defining y_{t+1}, the series to be forecast, and X_t, the N-dimensional series of predictor variables. These series are observed for time $t = 1, ..., T$. It is assumed that y and X have mean zero. The model with r^* common dynamic factors f_t is shown in the next equations,

$$y_{t+1} = \beta(L)f_t + \gamma(L)y_t + \epsilon_{t+1} \tag{4.1}$$

$$X_{it} = \lambda_i(L)f_t + e_{it}, \tag{4.2}$$

for $i = 1, ..., N$, where $e_t = [e_{1t}, ...e_{Nt}]'$ is a N x 1 idiosyncratic term and $\lambda_i(L)$, $\beta(L)$ and $\gamma(L)$ are lag polynomials in non-negative powers of L. It is assumed that $E(\epsilon_{t+1}|f_t, y_t, X_t, f_{t-1}, y_{t-1}, Xt-1, ...) = 0$.

The authors assumes that $\lambda_i(L)$, $\beta(L)$ and $\gamma(L)$ are modeled as having finite

orders of at most q. Explicitly $\lambda_i = \sum_{j=0}^{q} \lambda_{ij} L^j$ and $\beta(L) = \sum_{j=0}^{q} \beta_j L^j$. With this assumption it is possible to rewrite (4.1) and (4.2) as

$$y_{t+1} = \beta' F_t + \gamma L y_t + \epsilon_{t+1} \tag{4.3}$$

$$X_t = \Lambda F_t + e_t, \tag{4.4}$$

where $F_t = [f'_t, ..., f'_{t-q}]$ is r x 1, with $r \leqslant (q+1)r^*$, the i-th row of Λ in (2.4) is $[\lambda_{i0}, ..., \lambda_{i1}]$, forming a $N \times r$ matrix , and $\beta = [\beta_0, ..., \beta_q]'$. Here, $r*$ indicates the number of dynamic factors (f_t), and r represents the number of static factors (F_t). The assumption of finite lags for the factors means that the true number of factors underlying the dataset is finite, hence they can be gathered in a vector. Thanks to this static representation (which is a notational artifact the allows us to write the model in terms of static factors), it is possible to use principal component (PC) estimation. Of course, it is important to remember that this model would be inconsistent with infinite distributed lags of the common factors. Under a set of asymptotic rank conditions on Λ and moments conditions, the model allows for serial correlation of e_{it}. The moments conditions are:

(a) $\mathrm{E}(\mathrm{e}'_t e_{t+u}/N) = \gamma_{N,t}(u)$
$lim_{N\to\infty} sup \sum_{u=-\infty}^{\infty} |\gamma_{N,t}(u)| < \infty,$

(b)$E(e_{it}e_{jt}) = \tau_{ij,t}, lim_{N\to\infty} sup_t N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} |\tau_{ij,t}| < \infty,$

(c)$lim_{N\to\infty} sup_{t,s} N^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} |cov(e_{is}.e_{it}, e_{js}, e_{jt})| < \infty.$

Assumption (a) allows for serial correlation in the e_{it}, while assumption (b) allows for weak correlation across series. Assumption (c) limits the size of the fourth moment.

The PC estimator can be derived as the solution to the least square problem

$$min_{F_1,\cdots,F_T} V_t(\Lambda, F) = \frac{1}{NT} \sum_{t=1}^{T} (X_t - \Lambda F_t)'(X_t - \Lambda F_t),$$

subject to $N^{-1}\Lambda'\Lambda = I_r$. Solving this minimization problem, we obtain the estimate of the loadings and of the factors. The loadings estimates, $\hat{\Lambda}$, are equal to the eigenvectors corresponding to the r largest eigenvalues of matrix $X'X$, and the factors estimates $\hat{F}$ are given by $\hat{F} = \hat{F}(N^{-1}\hat{\Lambda}) = X'\hat{\Lambda}/N$.

In their paper, Stock and Watson specify various expectation maximization (EM) algorithms that allow to apply PC also in the case of missing observations and mixed-frequency data. It is interesting to report the adjustment of PC estimation in the case of one of the most common data irregularities, namely the presence of missing values. In this PC estimation the objective function is

$$V(F, \Lambda) = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it} - \lambda_i' F_t)^2. \tag{4.5}$$

In case we have an unbalanced dataset then (4.5) becomes

$$V^{\dagger}(F, \Lambda) = \sum_{i=1}^{N} \sum_{t=1}^{T} I_{it} (X_{it} - \lambda_i' F_t)^2, \tag{4.6}$$

where $I_{it} = 1$ if X_{it} is available and zero otherwise. To minimize (4.6) we need an iterative method, where the j-th step is defined as

$$Q(X^{\dagger}, \hat{F}, \hat{\Lambda}, F, \Lambda) = E_{\hat{F},\hat{\Lambda}}[V(F; \Lambda)|X^{\dagger}] \tag{4.7}$$

where $\hat{\Lambda}$ and $\hat{F}$ denote the estimates for Λ and F constructed in the j_{t-1}-st iteration. $X^{\dagger}$ is the full set of observed data and $E_{\hat{F},\hat{\Lambda}}[V(F, \Lambda)|X^{\dagger}]$ is the expected values of the complete data log-likelihood $V(F,)$, evaluated using conditional density of $X|X^{\dagger}$ at $\hat{F}$ and $\hat{\Lambda}$. The estimates of F and λ, at iteration j, solve $\text{Min}_{F,L} Q(X^{\dagger}, \hat{F}, \hat{\Lambda}, F, \Lambda)$. Equation(4.7) can be rewritten as

$$Q(X^{\dagger}, \hat{F}, \hat{\Lambda}, F, \Lambda) = \sum_{i} \sum_{t} \{E_{\hat{F},\hat{\Lambda}}(X_{it}^2|X^{\dagger}) + (\lambda_i' F_t)^2 - 2\hat{X}_{it}(\lambda_i' F_t)\}, \tag{4.8}$$

where $\hat{X}_{it} = E_{\hat{F},\hat{\Lambda}}(X_{it}^2|X^\dagger)$. The first term on the right hand side of (4.8) can be replaced by $\sum_i \sum_t \hat{X}_{it}^2$. If observations on X_{it} are missing, then at iteration j $\hat{X}_{it} = \hat{\lambda}_i' \hat{F}_t$. Estimates of F are updated by computing the eigenvector corresponding to the largest eigenvalues of $N^{-1} \sum_i \hat{X}_i \hat{X}_i$ where $\hat{X}_i$ is $\hat{X}_i = [\hat{X}_{i1}, ..., \hat{X}_{iT}]$. The estimates of Λ are updated by OLS regression of $\hat{X}^*$ on the updated estimates of F.

4.2 The Model by Forni, Hallin, Lippi and Reichlin (2000)

The idea of this model is again based on combining the approximate factor model and the basic dynamic factor model. The factors here are assumed to follow an infinite order moving average process (MA). The model can be represented by the following equation

$$y_{it} = b_{i1}(L)u_{1t} + b_{i2}(L)u_{2t} + ... + b_{iq}(L)u_{qt} + \xi_{it}, \tag{4.9}$$

here, L is the lag operator, the variable $\chi_{it} = y_{it} - \xi_{it}$ is called common component, or factor, and ξ_{it} is called idiosyncratic component of y_{it}. In this model, as for the approximate factor model, the cross-sectional dimension is tending to infinity. The model is completed by four assumption.

Assumption 1.
(I) The vector $[u_{1t}, u_{2t}, ..., u_{qt}]'$ is orthonormal white noise, i.e. $E(u_{jt}) = 0$, $var(u_{jt} = 1)$ for any j and t, $u_{jt} \perp u_{jt-k}$ for any j, t and $k \neq 0$, $u_{jt} \perp u_{st-k}$ for any $s \neq j, t$ and k. This means that $u_i t$ is serially-uncorrelated and not cross-correlated.
(II) The vector of idiosyncratic components $[\xi_{1t}, \xi_{2t}, ..., \xi_{nt}]'$, is a zero-mean stationary vector for any n, and $\xi_{it} \perp u_{jt-k}$ for any i, j, t and k. Notice here that the model is not assuming orthogonality of the idiosyncratic components.

(III)The filters $b_{ij}(L)$ are one-sided in L and their coefficients are square summable. As consequence of this assumption, the vector of variable $[y_{1t}, y_{2t}, ..., y_{nt}]'$, is zero-

mean and stationary for any n.

Before introducing the second assumption of the model it is useful to denote $\Sigma_n(\theta)$ as the spectral density matrix for the vector y_t, and $\sigma_{ij}(\theta)$ its entries.

Assumption 2
For any $i \in \mathbb{N}$, there exists a real $c_i > 0$ such that $\sigma_{ii}(\theta) \leqslant c_i$ for any $\theta \in [-\pi, \pi]$.

Denote λ_{nj} the function associating with any $\theta \in [-\pi, \pi]$, the real non-negative j-th eigenvalue of $\Sigma_n(\theta)$ in descending order of magnitude. This function is called dynamic eigenvalues of Σ_n. The dynamic eigenvalues of Σ_n^{χ} and of Σ_n^{ξ} are called common and idiosyncratic eigenvalues.

Assumption 3
The first idiosyncratic dynamic eigenvalue λ_{n1}^{ξ} is uniformly bounded, i.e., there is a real Λ such that $\lambda_{n1}^{\xi}(\theta) \leqslant \Lambda$ for any $\theta \in [-\pi, \pi]$ and any $n \in \mathbb{N}$.

Assumption 4
The first q common dynamic eigenvalues diverge in $[-\pi, \pi]$, i.e., $\lim_{n\to\infty}\lambda_{nj}^{\chi}(\theta) = \infty$ for j, in $[-\pi, \pi]$.

Assumption 3 and assumption 4 need more explanations. The first one simply allows for a limited amount of dynamic cross-correlation of y's. This assumption implies that the idiosyncratic causes of variation (ξ), have their effect on a finite number of cross-sectional units, tending to zero as i goes to infinity. Assumption 4 guarantees a minimum amount of cross-correlation between the common components. The model composed by equation (1) and assumptions 1 to 4, is called the generalized dynamic factor model.
It is interesting to see the definition of χ_{it}, in this model. First of all the authors remind, for a given spectral density matrix $\Sigma_n(\theta)$, the existence of n vectors of

complex valued functions

$$p_{nj}(\theta) = p_{nj,1}(\theta), p_{nj,2}(\theta), ..., p_{nj,n}(\theta),$$

$j = 1, 2, ..., n$, such that

(i) $p_{nj}(\theta)$ is row eigenvector of $\Sigma_n(\theta)$ corresponding to $\lambda_{nj}(\theta)$, i.e.

$$p_{nj}(\theta)\Sigma_n(\theta) = \lambda_{nj}(\theta)p_{nj}(\theta) \text{ for any } \theta \in [-\pi, \pi];$$

(ii) $|p_{nj}(\theta)|^2 = 1$ for any j and $\theta \in [-\pi, \pi]$;

(iii) $p_{nj}(\theta)\tilde{p}_{ns}(\theta) = 0$ for $j \neq s$ and $\theta \in [-\pi, \pi]$;

(iv) $p_{nj}(\theta)$ is measurable on $[-\pi, \pi]$;

where $\tilde{p}_{nj}(\theta)$ is the adjoint of $p_{nj}(\theta)$. A n-tuple of $p_{nj}(\theta)$, satisfying properties from (i) to (iv) is called dynamic eigenvector of $\Sigma_n(\theta)$. These dynamic eigenvectors can be expanded in Fourier series:

$$p_{nj}(\theta) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} [\int_{-\pi}^{\pi} p_{nj}(\theta)e^{ik\theta}d\theta]e^{-ik\theta}$$

Defining

$$\underline{p}_{nj}(L) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} [\int_{-\pi}^{\pi} p_{nj}(\theta)e^{ik\theta}d\theta]L^k,$$

then, for $j = 1, ..., n$, the scalar process $\{\underline{p}_{nj}(L)y_t, t \in \mathbb{Z}\}$, will be called the j-th dynamic principal component of y_t. Finally the authors define

$$\chi_{it,n} = \underline{K}_{ni}(L)y_t, \tag{4.10}$$

where

$$K_{ni}(\theta) = \tilde{p}_{n1,i}(\theta)p_{n1}(\theta) + ... + \tilde{p}_{nq,i}(\theta)p_{nq}(\theta)$$

Even though this model can be technically challenging, the main point to notice is that it relaxes the assumption of orthogonal idiosyncratic components. Another interesting feature of this model is that the estimation of the factors is based on the frequency domain, because $\underline{K}_{ni}(L)$ is function of the spectral density matrix $\Sigma_n(\theta)$.

Comparing this formulation to SW (2002b), this model does not require lag polynomials of finite order, but here the factor loading coefficients are not allowed to be time-varying. Another problem with this model, is that it requires two-sided smoothing, causing the estimates of the common components not available at the end of the sample. Forni and Lippi (2009) develop a one-sided estimator, which can be used for forecasting and other economic applications.

This overview of the factor models and its developments is intended as background for the empirical analysis, meaning that the core of the thesis lies in the empirical work, and this part is instrumental to understand the forecasting experiment. Even though I will use the model developed by SW, I wanted to illustrate other methods too, in order to give a general picture of the framework within which I am conducting my analysis. Developments for this work lie in using different estimation methods for the factors. Even though the methods developed by SW and by Forni et al. give similar factor estimates, there are new models that combine the maximum likelihood methods but they still allow the use of large datasets. These factor estimation methods are also known as "third generation" models, and it can be interesting to use them in forecasting experiments like the one of this thesis.

4.3 The Model by Kapetanios and Marcellino (2006)

The key factor of this model is that it retains the framework of a parametric state-space model, but it uses linear algebra methods based on subspace algorithms,

not maximum likelihood methods, to estimate the factors. This feature renders the model computationally feasible, even for very large datasets.

The model consists in the following state-space representation

$$x_{Nt} = Cf_t + \epsilon_t,\ t = 1, ..., T$$
$$f_t = Af_{t-1} + B^* v_{t-1}$$

where x_{Nt} is a N-dimensional vector of stationary, zero-mean variables at time t, f_t is a r-dimensional vector of factors at time t, and ϵ_t and v_t are mutually uncorrellated, standard orthogonal white noise sequences, of dimension N and r respectively. B^* is assumed to be non-singular. The main aim of this model is to estimate f_t, for $t = 1, ..., T$. The estimation of the factors is then based on singular value decomposition.

Even though this method has some advantages (namely, the parametric representation of the factors together with computationally feasible estimation), I stick to the SW method (2002b). This is due to two reasons: in Kapetanios and Marcellino (2006), the authors show that using this state space model gives comparable forecasting accuracy, compared to the static PC method of SW. The latter, is much easier to implement, so I will use it for the forecasting experiment. The second reason is that the parametric estimation method by Kapetanios and Marcellino (2006) does not readily handle datasets with a larger cross section compared to time dimension, which is the case for the StatFin and the micro dataset.

Part II

Empirical Analysis

In this part of the thesis, I conduct a forecasting experiment based on the three datasets mentioned in the introduction. After a quick look at the various empirical application developed in relation to factor models, I divide the empirical analysis in three sections which correspond to the different datasets.

5 Economic Applications

Dynamic factor models allow to include a large set of variables in the analysis, without suffering degrees of freedom problems. Statistical agencies and central banks collect a wide range of economic indicators, and these index models allow the use of this large amount of information in economic applications.
The main economic applications for dynamic factor models are:
(i) Construction of economic indicators.
The two most prominent examples of this application are the Chicago Fed National Activity Index, for US, and the EuroCOIN, for the Euro Area. An example of this application is found in Aruoba, Diebold, and Scotti (2009), where the authors use a dynamic factor model to create an index of economic activity, which is updated on weekly basis.

(ii)Forecasting.
This is the most common application for dynamic factor models. Many central banks and research institutions include estimated factors in the forecasting equation of the variable of interest. One seminal paper, where factor analysis is used in a forecasting environment, is by SW (2002a). In this study the authors show that forecasts made including a common factors are more successful in forecasting, compared to benchmark models like AR and VAR. Eickmeier and Ziegler (2008) conduct a comprehensive overview on the literature of dynamic factor models used in forecasting experiments.

(iii) International business cycles.

Dynamic factor models can be used to estimate the common driving force behind the economic performance of individual countries. For example, Breitung and Eickmeier (2005) apply this kind of analysis to the determination of the common factors behind European monetary union countries, and central and eastern europe countries. This kind of application requires the economic identification of the estimated factors, which is a rather difficult task.

(iv) Analysis of monetary policy.

Bernanke and Boivin (2003), use the SW method to estimate policy reaction functions and in Bernanke, Boivin and Eliasz (2005), the authors develop a factor augmented vector autoregressive model (FAVAR) to analyze the effect of monetary policy. Belviso and Milani (2003) estimate structural FAVAR to give an economic interpretation of the factors.

6 Empirical Analysis of the StatFin Dataset

6.1 Description of the Dataset

The dataset constructed using Statistics Finland data (StatFin) has three main sources: the StatFin dataset, the Bulletin of Statistics and the Astika database. From here on, with StatFin I mean the final dataset I constructed with Statistics Finland data. The StatFin database contains 144 monthly variables. These series start in September 2000 and end in July 2010. I cut some series to ensure to have a balanced dataset, which allows an easier estimation of the factors. These 144 variables include a range of financial indicators (e.g. Euribor rate, yields on bonds, OMX indexes), real economy indicators (e.g. turnovers, volume indexes, job vacancies) and price indicators (e.g. a number of producer price indexes). The dataset includes also various building permits and survey data (e.g. percentage of the interviewed sample who wants to buy home furnishing in the coming year). All the variables have been seasonally adjusted and log-differenced, if needed. Moreover, I standardize the whole dataset.

Of these 144 variables, I want to forecast four series, namely an indicator of GDP,

the industrial production (IP) indicator, the consumer price index (CPI) and the harmonized index of consumer prices (HICP). I consider the monthly growth rates of these variables. The forecasting period starts in July 2009 and ends in July 2010, meaning that I use almost nine years to estimate the various models and the remaining year of observations is left to be forecasted. I report next the plots of the series which are forecasted in this analysis.

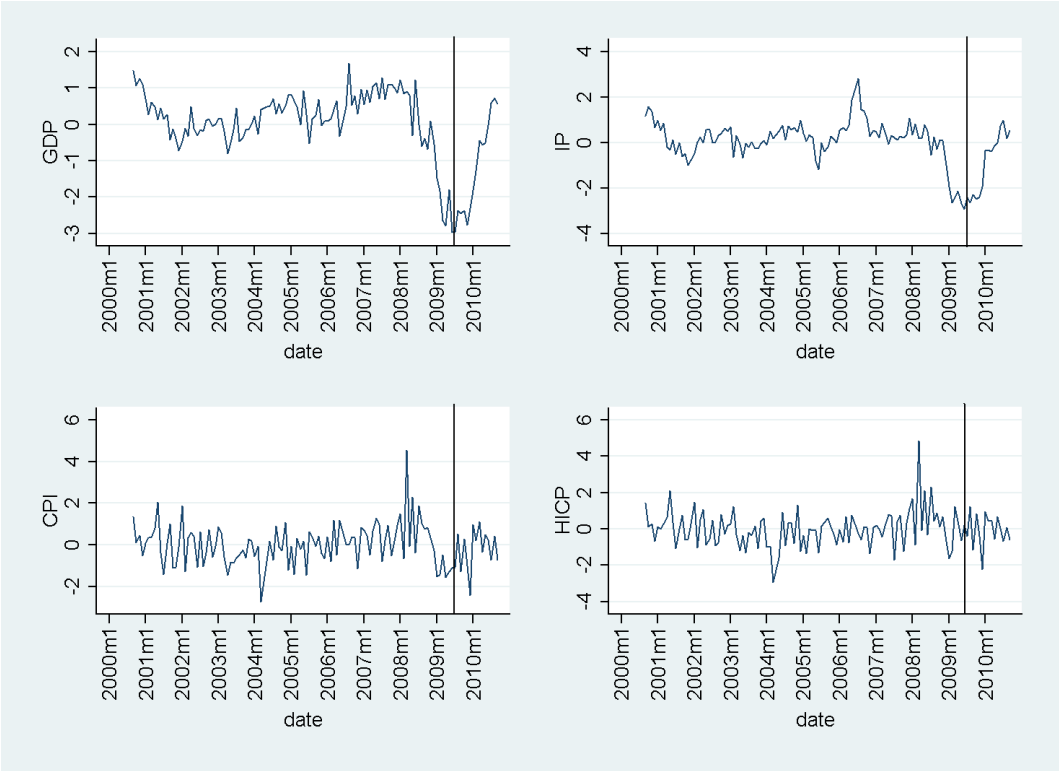

Figure 1: Variables to be Forecasted in the StaFin Dataset.

The vertical line in each graph indicates the beginning month of the forecasts. This period corresponds to the lowest point of the recent recession, right before the start of growth. This is mostly reflected in the graphs of GDP and IP. It is important to take into account the extremely peculiar period under consideration, when judging the forecasting performances.

6.2 Factors

The next step in the analysis is the extraction of the factors, using the SW estimation method. I estimate three factors from the StatFin datasets, excluding the variables that will be forecasted from the dataset. The choice of using only three factors is based on previous literature where the production of good forecasts is based on using only few factors. For example Stock and Watson (2002a) show that one or two factors are enough to obtain remarkable forecasting results. The plot of the three factors is reported next.

Figure 2: Factor Extracted from the StatFin Dataset.

These factors are estimated using the whole time span, therefore they are not the ones I use for forecasting purposes (to simulate a plausible forecasting environment). These plots are useful to give a hint on how we can interpret these factors. Factor 1 seems to carry information about the movement of the overall economy, where around mid-2008 the series starts to decline. This decline is followed by a growth starting in early 2009. If we compare the plot of factor 1 with the one GDP, it appears that the factor is lagging in the prediction of the start of the crisis, but it anticipates the restart of growth. Factor 2 and factor 3

graphs are not easily interpretable. The only intuition we can gather from these two graphs is the increase in the volatility of the dataset, starting in 2008. To try to give a more meaningful interpretation of the factors it might be useful to check the R^2 of the regression of each factor on each single variable included in the dataset. The tables reported next shows the ten largest R^2 obtained by these regressions. If $R^2 < 0.10$, then I do not report the variable and the related R^2.

Variable	Factor1	Variable	Factor2
6months Eur.	0.53	OMX Capitalization	0.42
3month Eur.	0.52	OMX Financial	0.34
1month Eur.	0.48	OMX Industrials	0.34
Reduced Time Worker	0.45	OMX Consumptions	0.33
Unemp.d Seeking Jobs	0.44	OMX Materials	0.25
Eonia Rate	0.40	Vol. Index Other Cons.	0.22
Turnover Adm.	0.40	OMX Energy	0.21
OMX Cons.	0.37	OMX Telecom	0.20
Import	0.36	OMX Cons. Staple	0.19
OMX industrials	0.35	OMX healthcare	0.19

Table 2: R^2 of the Regression of Dataset Variables on Estimated Factors.

Variable	Factor3
Turn. Textile	0.29
Vol. Index Chemicals	0.237
Turn. Food Industry	0.235
Turn. Forest Industry	0.22
Vol. Index Electronics	0.21
Vol. Index Man.	0.19
Vol. Index Investments	0.17
GDP	0.168
Purc. Price Inventory	0.16
Prod. Price Index Water	0.16

Table 3: R^2 of the Regression of Dataset Variables on Estimated Factors.

These tables show a very different picture from the one we could draw from the plots of the factors. Factor 1 appears to have a weak relation to GDP, IP and

other real economy variables. It seems that it relates mostly to interest rates like the Euribor and the EONIA rate. It also has a pretty large R^2 on unemployment related variables. From this R^2 analysis I cannot draw a clear interpretation for factor 1. Factor 2, instead, seems to be easily interpretable: it can be read as an indicator of the stock market. Finally, factor 3 is related mostly to variables we can consider as real economy indicator. It is worth noticing that none of the factors (except for factor 3) has any relation to price indicators, suggesting than the forecasting performance for CPI and HICP will be disappointing. This R^2 analysis allowed me to judge the utility of factors other than the first three. Additional factors did not add any relevant information.

One last feature of the factors, that it can be interesting to explore, is the how many factors are needed to explain the variance of the whole dataset. This can be seen by the plot of the sum of the greatest eigenvalues of the matrix $X'X$. I choose to plot the first 20 eigenvalues. If the resulting curve is very steep at the beginning, it means that the eigenvalues after the first few are small. This implies that factors corresponding to this small eigenvalues explain little variance of the dataset. I report this plot next.

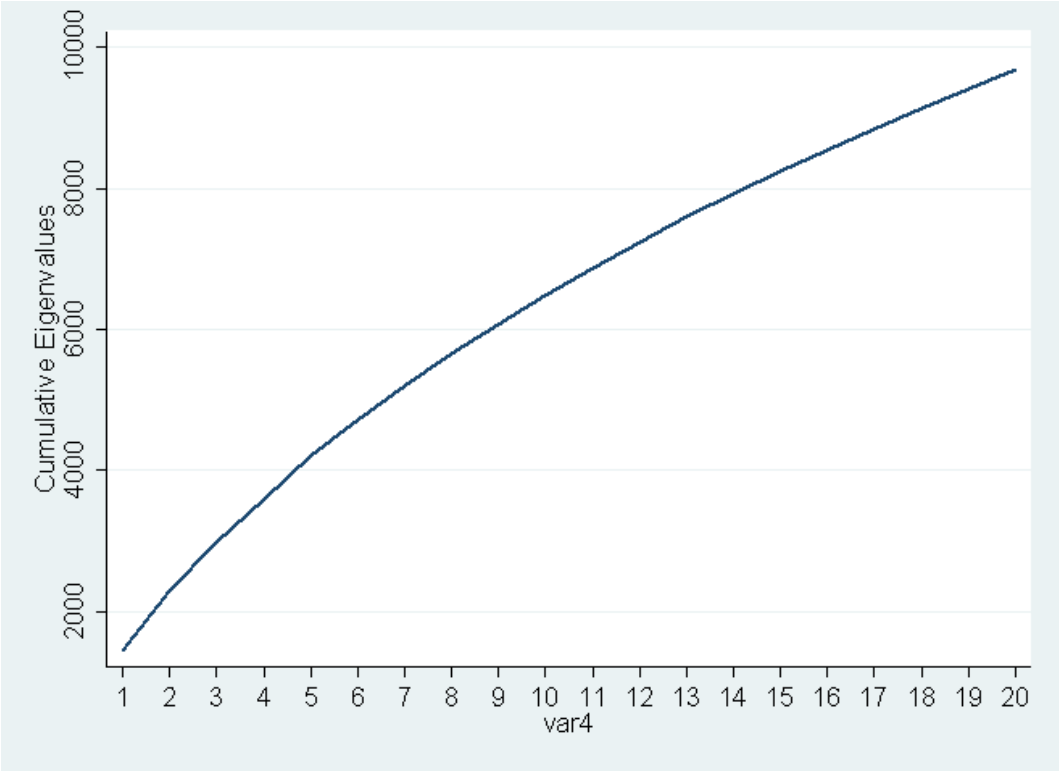

Figure 3: First 20 Eigenvalues, Cumulative

The sum of the first 20 eigenvalues shows that, to explain the variation of the StatFin dataset, many factors are needed. This is seen in the almost linear growth of the cumulative sum of the eigenvalues. It is important to underline that this graph does not give any information about the number of factors needed in the forecasting of the variables of interest. This is because this analysis suggests the number of factors needed to explain the variation of the whole dataset, not the movements of any particular variable.

6.3 Forecasting Results

For the StatFin dataset I will forecast the 12 months period starting in July 2009 and finishing in July 2010, for the GDP, IP, CPI, and HICP series. The evaluation the various forecasts is based on the relative mean square forecast errors ($MSFE$), where the benchmark model is an autoregressive model (AR) model, with lags order selected by using the Schwarz information criteria (BIC). I also estimate forecasts using a vector autoregressive model (VAR), where I use 3 variables. GDP or IP, CPI or HICP and the 3months Euribor rate (following the example of SW (2002a)). This formulation allows to have, in the VAR model, an economic activity indicator, a price indicator and a financial indicator.

The forecasts based on the factors follow two models: one is augmented with autoregressive terms, and the other follows the FAVAR model. In the first case, the factors are not interacting with the forecasting variable, while in the FAVAR model the factors have a close interrelation with the forecasted variable. To simulate a forecasting environment, in the factors plus AR term, I need to use the lags of the factors corresponding to the forecast horizon (which I indicate as h). For example, if the forecast horizon is six months, I have to use the 6-th lag of the factors. In the FAVAR this problem is not relevant because I let the model forecast new factors as it iterates for the forecast horizon. The equation for the factors plus AR model is:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \beta_1 f_{1,t-h} + \beta_2 f_{2,t-h} + \beta_3 f_{3,t-h}$$

The amount of autoregressive terms has been decided by BIC and by forecasting considerations. For the FAVAR, the forecast equations are given by:

$$\begin{aligned} y_t =& \phi_0 + \Phi L(2) y_t + \beta L(2) F_t \\ F_t =& \Psi_0 + \Gamma L(2) y_t + \Psi F_t \end{aligned}$$

Where F_t, Φ, β,Ψ_0,Γ and Ψ are 3×1 vectors. F_t contains the factors estimated, while y_t is the variable of interest in the forecast. L indicates the lag operator. Another point which is important to underline is that I use a different set of factors compared to the one I previously shown. For the forecasts, I extract the factors using the whole dataset except the part I want to predict, and I subsequently estimate recursively, month by month, factors until the last period of the forecast. I also include a model that instead of using the factors lagged by the forecasting horizon, it uses always the 12-th lag of the factors. This model is equal to the factors+AR model, for the one year ahead forecasts, while it creates different forecasts for the other horizons. This model is represented by:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \beta_1 f_{1,t-12} + \beta_2 f_{2,t-12} + \beta_3 f_{3,t-12}$$

The results of the forecasts, in terms of relative $MSFE$, are reported next. If the relative $MSFE$ is larger than one it means that the model under consideration creates less accurate forecasts, with respect to the AR model. The converse is true if the relative $MSFE$ is lower than one.

Variable	Horizon	AR	VAR	Fac+AR	FAVAR	Fac(12)+AR
GDP	Dyn	1	0.74	0.10	0.30	0.10
	h=6	1	0.59	0.47	0.22	0.15
	h=3	1	0.62	0.94	0.31	0.26
	h=1	1	0.82	0.87	1.03	0.36
IP	Dyn	1	0.85	0.41	0.29	0.41
	h=6	1	0.80	0.38	0.22	0.45
	h=3	1	0.79	0.94	0.41	0.69
	h=1	1	0.928	0.84	0.85	0.87

Table 4: Relative $MSFE$ for Economic Activity Indicator Forecasts.

One word about the forecast horizons. Because of the short forecast period (1 year), the longer h-step ahead forecasts cannot be interpreted in the usual way. The longest forecasts I have are the one year dynamic forecasts, which consist in the prediction of the period going from July 2009 to July 2010, using only the sample up to June 2009. Usual 12-th steps ahead forecasts correspond to the one year later predictions computed month by month, while here the only real one-year later forecast is the last one. For shorter h, this problem is reduced. However, this issue concerns all the forecasts, hence they are still comparable. The main reason why I use such a short forecast period is that the theory on dynamic factor models concerns asymptotic properties. The small sample properties of factor estimation methods (for example the model by SW, 2002b) have not been investigated. It is an interesting theoretical line of research to see how factor estimation behaves in the presence of relatively short datasets. Because of this, instead of using the classical 12-steps forecasts, I use the one year dynamic forecasts. Formally, these forecasts generate the following MSFE :

$$\sum_{h=1}^{12}(\hat{y}_{t+h} - y_{t+h})^2,$$

where t is July 2009. Usual 12-months ahead forecasts are given by:

$$\sum_{t=s}^{e}(\hat{y}_{t+12} - y_{t+12})^2,$$

where s is the last observation of the in-sample and e is the last observation for which the 12-months forecasts are computed. The first 12 predicted observations, for 12-steps ahead forecasts, are dynamic forecasts starting at the beginning of the out-sample, while the rest are computed month by month, with one year horizon. This problem is present for this dataset and for the micro dataset, while the BOF dataset allows me to create standard 12-step ahead forecasts. The short length of the forecasting period, and its peculiarity, must be kept in mind while evaluating these forecasting performances.

The results are very encouraging. Both for GDP and IP, the models based on the factors create better forecasts compared to the AR and the VAR model. For example the MSFE obtained by using the 12 lag factor model is only 10% of the one obtained by the AR model. Another surprising result is that using the 12 lag model for GDP, which implies using less information than in the usual h lag model, creates considerably better forecasts. This might be the sign of leading relationship between the factors and GDP and IP. Of course, these results must be taken with caution because the period of the forecast is very peculiar, which renders basic models like the AR very weak. Parallel to this analysis, I tried to forecast two years of observations, instead of just one. This allowed me to make proper 12-steps ahead forecasts. The performance of the factors-based model was not satisfactory, never beating the AR benchmark. I believe that this was due to the lack of observations in the estimation of the factors. The study of small-sample properties of factor estimation is a research possibility that can stem from this work. Next, I report the plot of the one year ahead forecast against the original series.

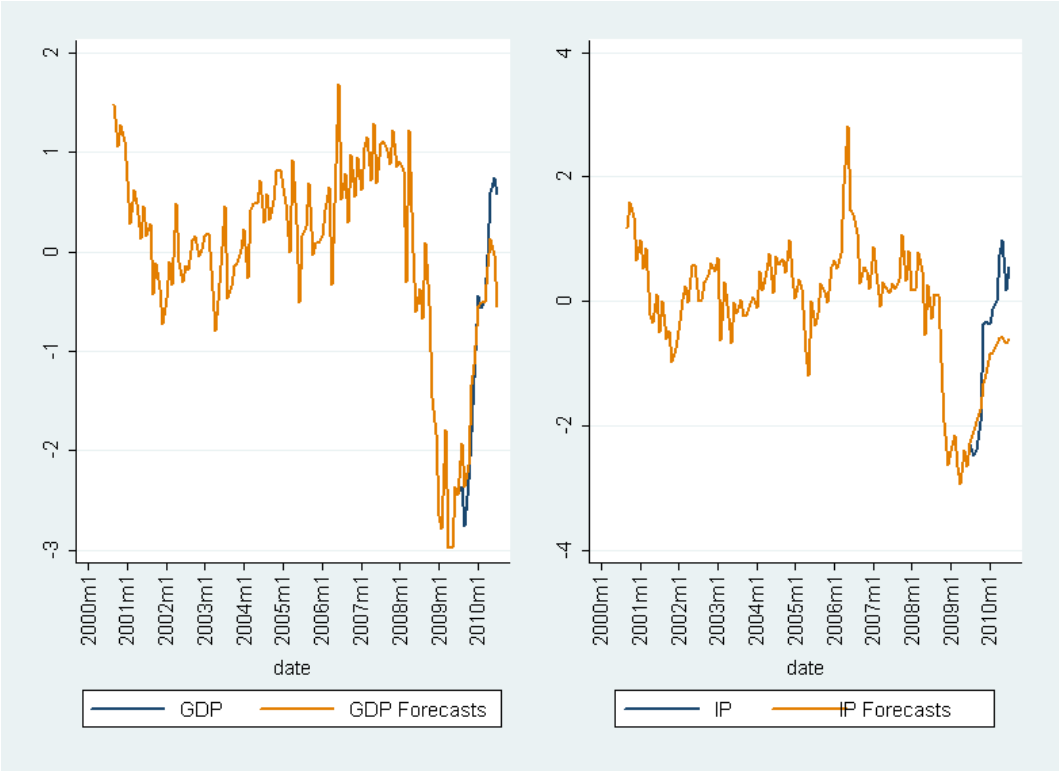

Figure 4: One Year Dynamic Forecasts for GDP and IP using Factors + AR Method

The plot shows the ability to predict the one year period we consider in this analysis, by the factor plus autoregressive term model. The orange line indicates the forecasts, which overlaps perfectly the original series before July 2009. The fit is surprisingly good for the GDP series, while it seems that the IP forecasts underestimate the value of interest. Both forecast series underestimate the last observations in the period under consideration. Still, they manage to capture the small decline in the end of the forecasting period.

I replicate the same forecasting analysis for the CPI and HICP series. The table of the relative $MSFE$ for the forecasts of the price series is reported next.

Variable	Horizon	AR	VAR	Fac+AR	FAVAR	Fac(12)+AR
CPI	Dyn	1	1.50	0.49	1.08	0.49
	h=6	1	1.45	0.34	1.07	4.96
	h=3	1	1.20	0.70	1.02	4.14
	h=1	1	1.14	0.87	1.32	2.44
HICP	Dyn	1	1.17	1.66	1.27	1.668
	h=6	1	1.18	0.64	1.29	4.29
	h=3	1	1.21	0.78	1.022	3.53
	h=1	1	1.27	1.17	1.651	2.31

Table 5: Relative $MSFE$ for Price Indicators.

The results are different from the previous one. It seems that the factors contain less information for the price series than for the economic activity, even though the Fac+AR model gives significantly better result compared to the other models. Here the 12 lag model performs worse than the one with h lags. It might be a sign of a shorter term relationship between the factors and the price series. It is very surprising the finding that the h-lag factor plus AR model gives very good forecasts for all the forecast horizons. Literature on dynamic factor models report a very pessimistic view about the ability of factor models to forecast price series. This sharp contrast with common findings reminds us that the period we are considering in the forecasting exercise is particular, to say the least. All these results must be taken we caution. The plots in the next page represent the forecasts of the price indicators against the original series.

Figure 5: One Year Dynamic Forecasts for CPI and HICP using Factors + AR Method

As suggested by the $MSFE$s, the plot signals a good ability by the factor models to forecast the price series. In particular, it is interesting to notice that the forecast series is able to reproduce the high volatility of the original price series. Even though the forecasts approximately mimic the behavior of the original series, the last observations are very different. Toward the end, the forecast series and the actual series seem to diverge. This overall good performance of the dynamic factor model to forecast price series is in contrast with the usual findings in the literature, as showed by SW(2010).

7 Empirical Analysis of the BOF Dataset

7.1 Description of the Dataset

The BOF dataset is considerably longer than the StatFin one. It starts in September 1987 and ends September 2010, giving 12 years of data. This time span includes two important recessions (the recent one and the Finnish banking crisis of early 1990's), which can be challenging for the estimation of the factors and the forecasts. The series are again monthly. The greater length of this dataset is the main reason why it is examined in this paper. First of all, it allows me to compute classical 12-steps ahead forecasts, giving a more reliable evaluation of the forecasting performance. The fact that the time span includes two important recession is another features that renders the dataset, desirable to use.

The number of variables here is lower than in the StatFin case, being 104 variables including the ones I want to forecast. The range of variables is still ample, containing a wide range financial variables and real economic indicators. The dataset, however, lacks some measures that were included in the StatFin one, for example building permits for a range of construction types or OMX indexes for different sectors. Moreover, the price series of the StatFin dataset were more complete with respect to the ones included in the BOF dataset. Comparing the forecast results for these different datasets may allow to shed some light on the conditions for reliable factor estimations and for good forecasts. As before, the series have been seasonally adjusted and log differenced, if needed.

The variables I am interested in forecasting are the indicator of GDP, the industrial production series (IP) and the consumer price index (CPI), both yearly changes and month-to-month changes. The forecasting period starts in 1998 (January) and ends in September 2010. This longer forecasting time allows to have better assessment of the predicting power of the factors-based models. Moreover, the forecasts for the 12 months horizon follow the traditional definition. Below, I report the plots of the variables to be forecasted.

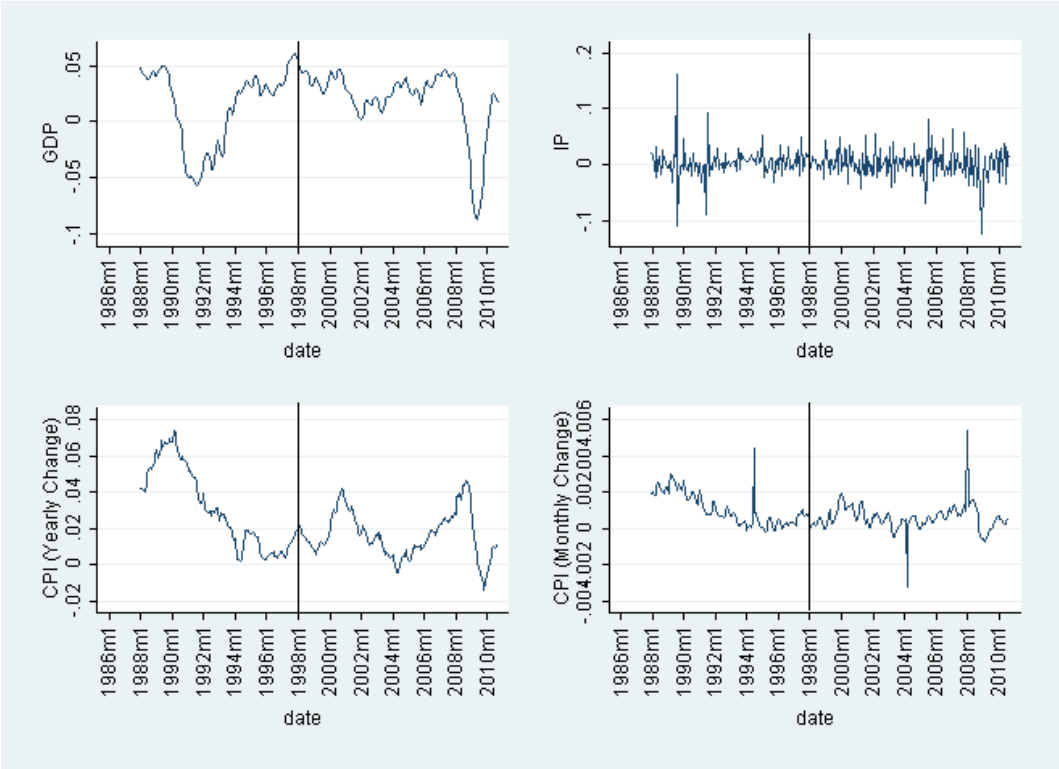

Figure 6: Variables to be Forecasted in the BOF Dataset

As before, the vertical lines indicate the beginning of the forecasting period. It is worth noticing that this series are different from the one in the StatFin dataset. The GDP indicator, for example, looks much smoother than the series present in the StatFin dataset.

7.2 Factors

I extract the factors through the PC methodology, using the dataset without the variables to be forecasted. As before, I estimate the factors using the whole time span, while I use the previously described recursive estimation for the forecasts. The use of the whole dataset is motivated by the fact that I want to take a descriptive look over the factors, to try to give them an economic meaning, without considering the real environment in which the forecaster operates. The factors plots are reported below.

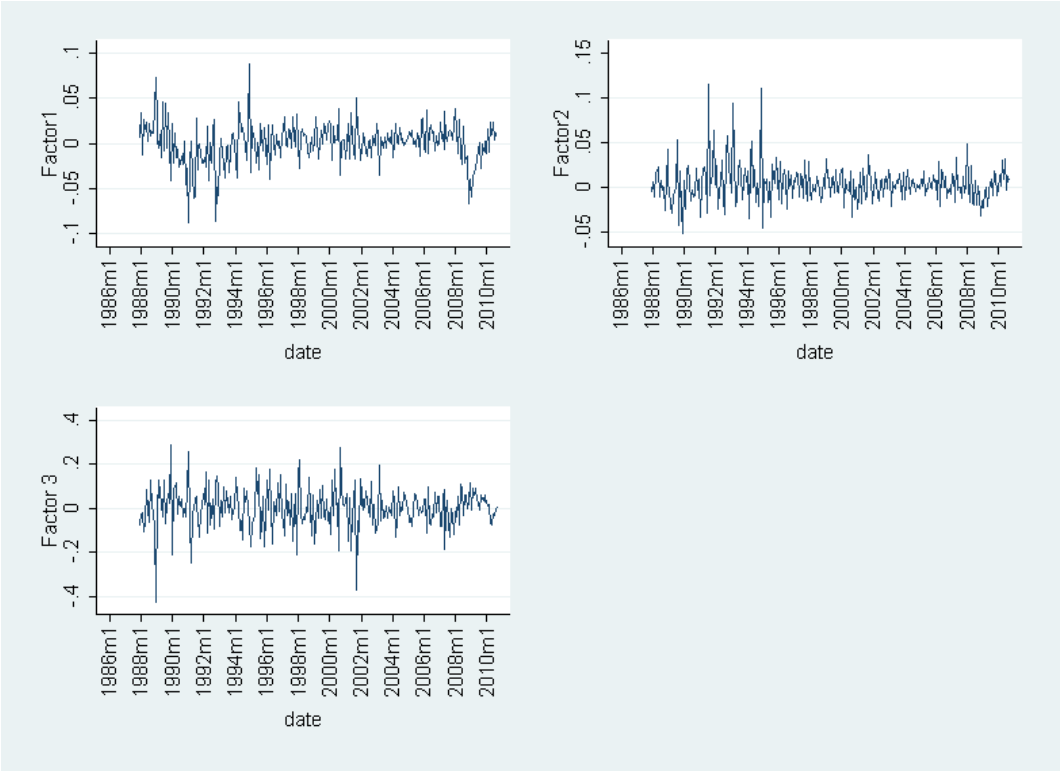

Figure 7: Factors Extracted from the BOF Dataset

These plots do not give easily interpretable information. The only clearly noticeable characteristic is found in the plot of factor 1. It seems that the factor manages to capture the recent crisis pretty well, even though the recession of the early 90's is not well represented. Factor 1 does drop around the beginning of the 90's, but the decrease is not as well defined as for the recent financial crisis. All the factors are very volatile compared to the series forecasted in this analysis. To try to give an interpretation to the factors I replicate the R^2 method used for the StatFin dataset. In the next page, I report the tables containing the first ten largest R^2.

Variable	Factor1	Variable	Factor2
M1 Annual Growth	0.40	C.A. Exp. on Goods and Serv.	0.53
C.A. Exp. on Goods and Serv.	0.26	C.A. Exp. on Goods	0.52
C.A. Exp. on Goods	0.26	Value of Import Goods	0.48
Value of Import Goods	0.25	Trade Value Imp. EU	0.46
M3 Annual Growth	0.44	Value Imports EU	0.43
Unemp. Job Seekers	0.24	Total Vol. Good Imports	0.41
Members Unemp. Funds	0.24	C.A. Factor Payments	0.35
Trade Value Imports EU	0.23	C.A Exp.	0.32
Total Vol. Good Imports	0.22	Imports non-EU	0.30
GDP	0.21	Value Exports	0.27

Table 6: R^2 of the Regression of Dataset Variables on Estimated Factors.

Variable	Factor3
M2 Annual Growth	0.87
M3 Annual Growth	0.72
M2 Stem	0.35
M3 Stem	0.35

Table 7: R^2 of the Regression of Dataset Variables on Estimated Factors.

The tables of R^2s do not allow us to assign a well defined meaning to the factors, but it is still possible to get some insights. First of all, none of the factors seems to have relation to the price series, suggesting poor forecasting ability for the CPI series (this element is probably due to the lack of price indicators in the dataset). On the other hand, most of the variables which are related to factors are nominal, except few (e.g. the number of job seekers, volume of imports). The only factor which is interpretable in a rather straightforward way is factor 3. This factor looks like being a pretty good indicator of the monetary aggregates. Factor 1, instead, is too heterogeneous to get a clear picture. Factor 2 has a tendency to effect trade variables (mostly imports), but this relation is not univocal.

I report, as before, the first 20 cumulative eigenvalues for the BOF dataset, to see the relation between the factors and the total variance of the dataset.

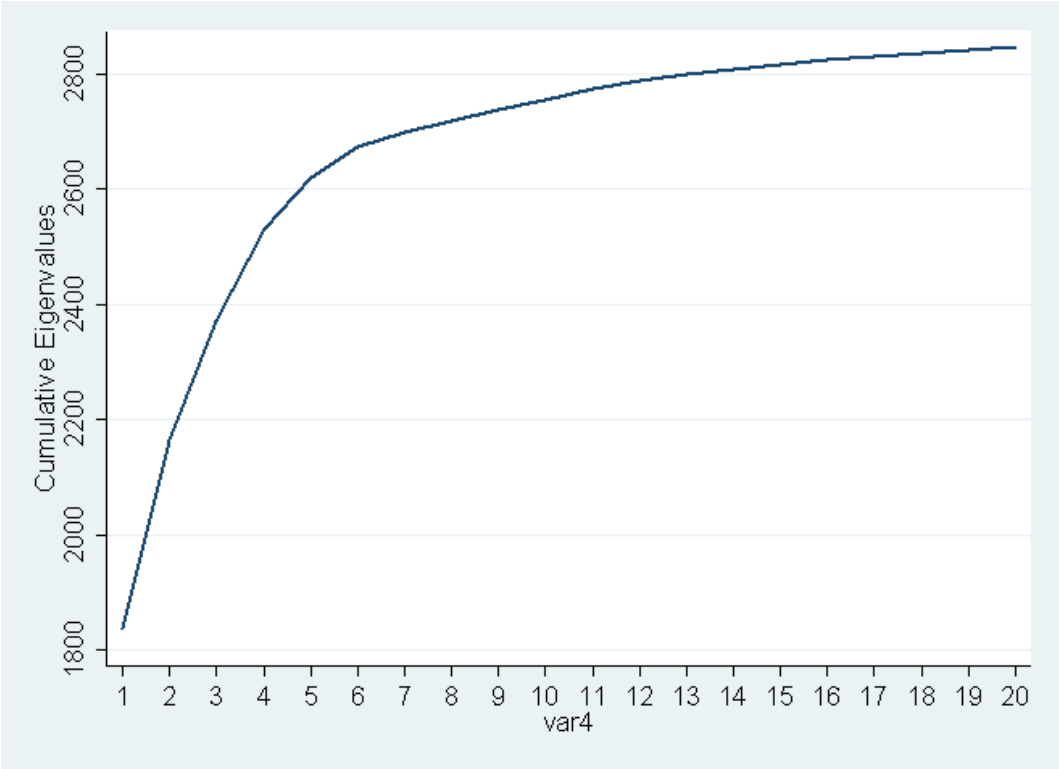

Figure 8: First 20 Eigenvalues, Cumulative

The picture is radically different compared to the one obtained in the StatFin dataset. The increase at the beginning is much steeper. From the 5th eigenvalue onward, the curve is very flat, indicating that the eigenvalues after the first five contribute very little to the variance of the dataset. This may indicate a lower range of variables, compared to the StatFin dataset. It is true that this dataset lacked many building permits indicators and survey data on households consumption plans.

7.3 Forecasting Results

As mentioned before, the forecasting period starts in January 1998 and ends in September 2010. The estimation of the factors and of the forecasting models starts in september 1987 and ends in January 1998. After that date, the factors are estimated recursively month by month, to simulate the forecaster's environment. The series of interest are GDP, IP and CPI (monthly and yearly changes).

I formulate six forecasting models. The first five are similar to the ones introduced before: an AR model, a VAR where I use the economic activity indicator (GDP or IP) together with the CPI and the 3-months interest rate, the FAVAR model, the 12-lag factors plus AR component model and the h-lag factors plus AR model. The last model is a modification of the VAR model where, instead of the 3-months interest rate, the first factor is included. The reason behind this choice is that the dataset lacked a comprehensive number of price variables. My guess is that the factors could not capture at all the price movements, which could be useful for the economic activity forecasts.

The comparison of the various forecasts is still based on the relative $MSFE$ and the results are reported in the next table.

Variable	Horizon	AR	VAR	Fac+AR
GDP	h=12	1	0.69	0.98
	h=6	1	0.80	1.015
	h=3	1	0.88	1.002
	h=1	1	0.92	1.003
IP	h=12	1	0.994	1.009
	h=6	1	0.98	1.007
	h=3	1	0.99	1.03
	h=1	1	0.95	0.94
Variable	Horizon	FAVAR	Fac(12)+AR	VAR(Fac1)
GDP	h=12	1.02	0.98	0.64
	h=6	1.04	0.98	0.83
	h=3	1.07	1.00	0.97
	h=1	1.09	0.9997	1.08
IP	h=12	0.999	1.009	0.996
	h=6	1.002	1.009	1.62
	h=3	1.019	1.009	1.004
	h=1	0.95	1.003	0.94

Table 8: Relative $MSFE$ for Economic Activity Indicator Forecasts.

From this table it is clear that the factor models here specified do not provide good forecasts for this dataset. The factor models are usually outperformed by the VAR model. The notable exception to this finding is the 12 months ahead forecasts of the VAR model with factor 1 instead of the interest rate variable. This forecast produces a decrease in the MSFE of 5% with respect to the VAR, which is a significant improvement. Also the 1 month ahead forecasts for IP are better for factor models, with respect to the VAR.

Even though the results are disappointing, there are some reasons I believe can justify the scarce performance of these class of models. First of all, the variable range might be too narrow. To ensure to have a longer dataset, I had to cut some interesting indicators which covered short time span. Another problem that forced me to give up many variables was the presence of outliers. Many series showed a considerable number of extreme observations, which heavily affected the estimation of the factors. To create reliable factors I had to drop all the variables showing too many outliers. It could be interesting to investigate this point, developing simulation studies on the effect of outliers on factors estimation. Another point that the reader must bear in mind is that the estimation method for the factors is a very simple one. Although Boivin and Ng (2005) and D'Agostino and Giannone (2006) showed that PC estimators and the Forni et. al. (2005) method give factors that produce similar forecasts, it would be interesting to extract the factors using so called "third generation" methods. These methods, described for example in Doz, Giannone, and Reichlin (2006), could produce substantial improvements in the estimation of the factors. One last improvement can be obtained by creating more complex forecasting models, including more factors and more lags. I do not go in this direction, in the current work, because I want to use the simplest model possible. I report next the table with relative $MSFE$ of the various forecasting models for the two CPI series.

Variable	Horizon	AR	VAR	Fac+AR
CPI(Yearly)	h=12	1	0.84	1.008
	h=6	1	0.69	1.05
	h=3	1	0.71	0.93
	h=1	1	0.84	0.99
CPI(Monthly)	h=12	1	1.04	1.08
	h=6	1	0.93	0.97
	h=3	1	0.99	1.03
	h=1	1	0.9801	0.983
Variable	Horizon	FAVAR	Fac(12)+AR	VAR(Fac1)
CPI(Yearly)	h=12	0.99	1.008	0.83
	h=6	0.95	1.037	0.74
	h=3	0.93	1.036	0.77
	h=1	0.98	1.02	0.89
CPI(Monthly)	h=12	1.01	1.08	0.95
	h=6	0.92	0.97	0.95
	h=3	1.003	1.039	1.032
	h=1	0.988	1.003	1.01

Table 9: Relative $MSFE$ for CPI (Yearly and Monthly Changes) Forecasts.

The relative $MSFE$s show the difficulty for the factor models to forecast the price series. For the yearly change in CPI, only the 12-steps ahead forecasts given by the VAR model including factor 1 are better than the ones of the basic VAR model. It is also important to notice that the FAVAR model produces better forecast with respect to the AR model, even though the improvements are not as dramatic as before. The main reason for this unsatisfactory performance is the small amount of price series in the dataset. Because of this reason, the estimated factors do not capture the relevant information that are needed to forecast the CPI. These results change only slightly if we concentrate on month-to-month changes in CPI. Here the VAR benchmark is beaten twice by factors-based models. It is also important to notice that the AR model perform better than for the

yearly changes in the same variable. Also, the forecasts computed for the CPI monthly changes series present a worse MSFE compared to the ones in the StatFin dataset for the same series. I believe that this is due from the fact that the period to be forecasted here is much longer. Another element that can justify these disappointing results, is that the BOF dataset contained fewer variables than the StatFin one, which can cause a worse estimation of the factors. The plot of the forecasts for the BOF dataset is reported in the next page.

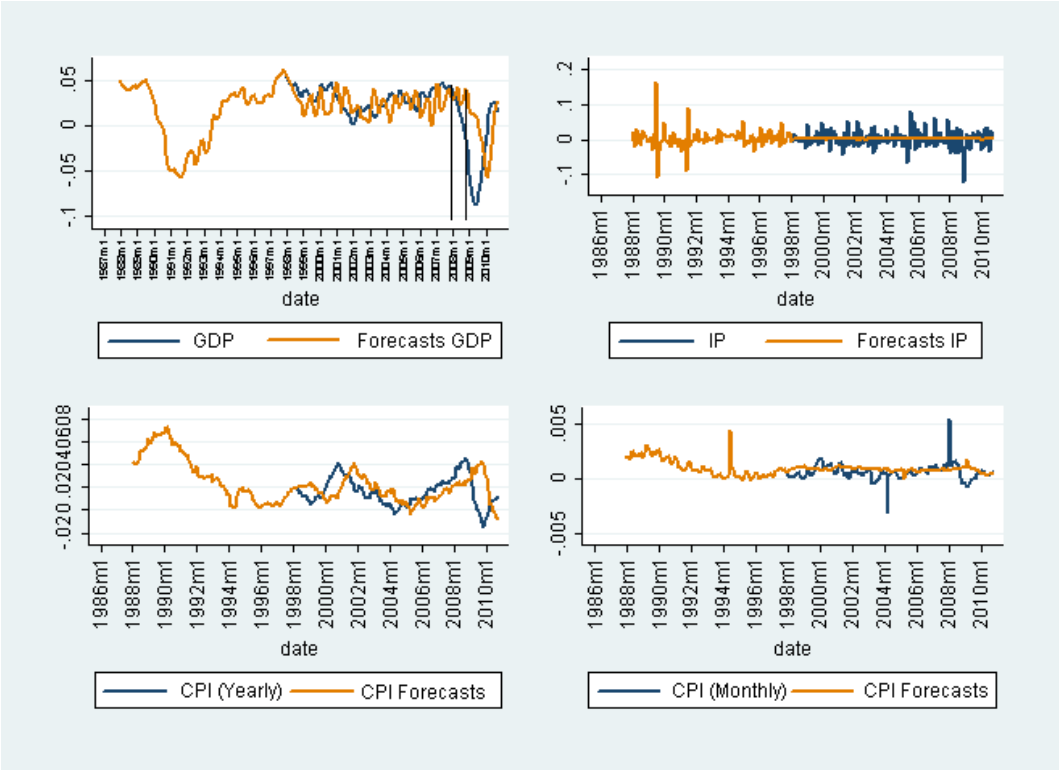

Figure 9: 12-Steps Ahead Forecasts for GDP, IP and CPI (Yearly and Monthly Changes) Using the Factors+AR Model

The first feature I want to underline from these plots is the behavior of the IP forecast. This series has a much lower volatility with respect to the original IP series. This problem does not depend on using a factor model, because also the AR model and VAR model give very similar results. The forecasts for IP do not give useful information about the behavior of the series. The CPI (year change) forecasts, instead, capture the magnitude of the variation of the series but, for many periods, the forecast shows a clear lagging relationship compared to the original series. This means that the forecasts need a quite considerable amount of

time to detect the movement in the series. This is not true, however, for the period starting around 2005 and ending in the beginning of 2008. During this time span, the forecasts are fairly close the original series. The series of monthly changes of CPI, presents more difficulties in forecasting. The factors-based forecasts do not manage to capture precisely the variability of the original series. The most interesting plot is the one of GDP. First of all, the forecasts generated by this model are more volatile compared to the GDP series. Another interesting characteristic is how the model is fairly good in forecasting the period between 2002 until 2008, a period of low variability. From 2008 onward the forecasts take more time to predict the movements in GDP, mostly for the sharp decline of the recent recession. Such behavior can be explained by the heavy structural break due to the recession. SW (2009) focus on the forecasting ability of factor models in case of a single break. They find out that the factors remain well estimated, but that it is important to incorporate the instability in the forecasting equations. A positive feature of these factors-based forecasts, is the ability to predict the restart of growth. Even though there is a considerable lag at the beginning of this period, the forecast and the original series converge quite quickly. One last point I want to highlight is the time the model needs to predict the start of the recent recession. The two vertical lines I draw on the plot signal the start of the recession, as indicated by the two series. From a closer look, it appears that the forecast series predicts the start of the recession around ten months later than the actual beginning. Statistics Finland estimates were able to predict the start of the recession one year later than it actually begun. The result for this analysis shows a slightly better performance of the factor models.

The analysis of the BOF dataset has given less positive signals about the ability of the factor models to forecast macroeconomic series. These results are in sharp contrast to the StatFin dataset ones and are different from many forecasting experiments based on US data. For example Boivin and Ng (2005) show that the $MSFE$ for factor forecasts, compared to the AR benchmark, ranges between 055 to 0.83 at 6-month horizon and 0.49 to 0.88 for 12-month horizon. These results hold for real series of US (IP, employment, real manufacturing), while the $MSFE$ is around 0.9 for inflation series. I believe that the main reason for

the disappointing performance of the factor models stands in the deficiencies of economic indicators in the BOF dataset.

8 Empirical Analysis of the Micro Dataset

8.1 Description of the Dataset

This dataset has been constructed on data provided by Statistics Finland. It consists in 527 variables, with time span starting in February 2000 and ending in November 2010. The variable range includes a set of turnovers and wages for industry, aggregated at 3-digits levels, resulting in 138 turnover series and 115 wage series. The initial turnover and wage series were at the firm level, but, to have a dataset which is easier to handle, I perform a low level aggregation. The data also include 148 price series, which indicate prices for different products and for different regions of Finland. Finally, the dataset contains 167 building permits variables. These consist in monthly permits, indicated as volume, for different types of buildings and different Finnish regions. Again, the variables have been seasonally adjusted and log differenced, in case of the presence of unit root.
The use of a dataset based on micro variables is one of the main contributions of this thesis, together with applying dynamic factor models to a large Finnish dataset. It is interesting to check the ability of this dataset to forecast the macroeconomic variable of interest of the StatFin dataset, namely GDP, IP, CPI and HICP. The micro dataset contains a very large number of variables, but the type range of these variables is pretty narrow. The dataset does not contain any financial variable, trade variables and other macroeconomic indicators that could add information. My intuition is that this dataset should perform worse in forecasting the previously cited variables, compared to the StatFin results. One important application that can be derived by this kind of dataset is the possibility of estimating missing values of one of the variables of the micro dataset. If, for example, some observations for a price series are missing, we can estimate factors from the rest of the dataset, and use the estimated factors to "nowcast" the missing values. Another interesting use of this model in relation with micro data, can be found in the creation of flash estimates of GDP. Micro data are quicker

to obtain, for statistical agencies, with respect to aggregate measures like GDP. Dynamic factor models can be a very fast procedure to get initial estimates of aggregated indicators, basing the factor estimation on the data available. This practice is called "nowcasting" and some important papers describing this method are Giannone, Reichlin and Small (2005), and Angelini, Camba-Méndez, Giannone, Rünstler and Reichlin (2008).

As before, I create 12, 6, 3 and 1-step ahead forecasts, starting in July 2009 and ending in July 2010, for the same variables I forecasted in the StatFin analysis.

8.2 Factors

As for the other two datasets, I use the SW method to extract the factors. The factors here represented are based on the whole dataset and they are not the ones used for the forecasting exercise. These factors are based on the micro dataset, including observations that go from September 2000 to July 2010, to ensure compatibility with the StatFin dataset. Below I report the plot of the three factors extracted.

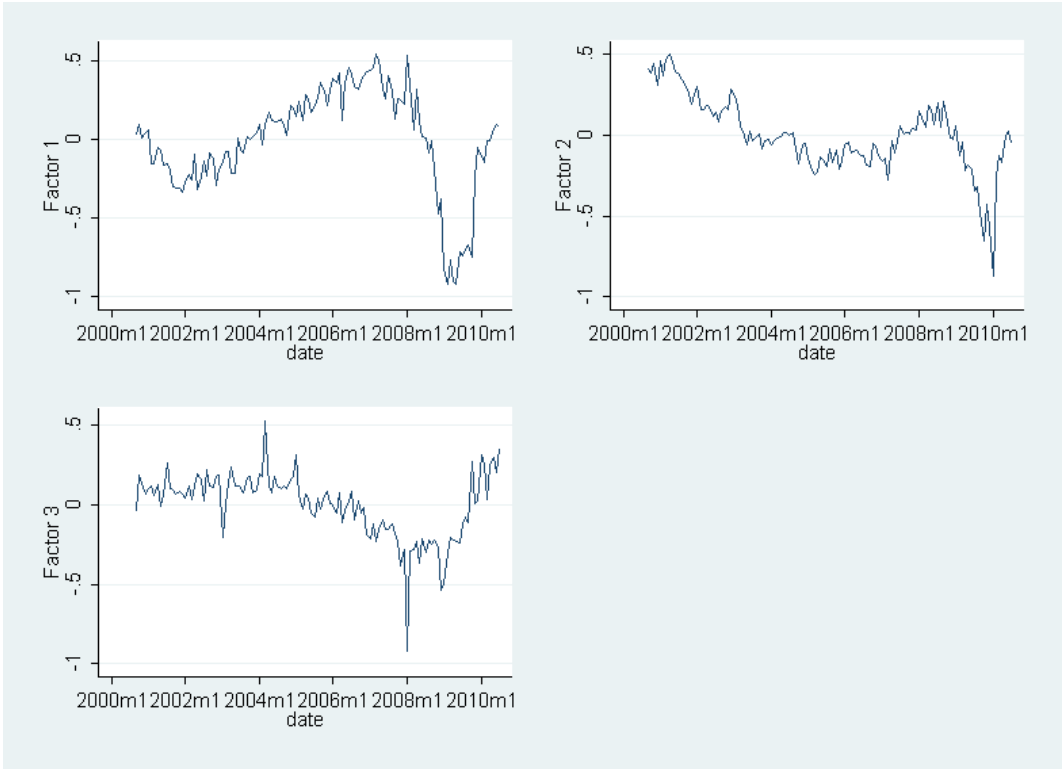

Figure 10: Factors Extracted from the Micro Dataset

The first surprising feature of these factors is the smoothness. Compared to the factors extracted from the other two datasets, factor 1 and factor 2 are much less volatile, describing a clear trend. For example, factor 1 seems to represent well the overall Finnish economic activity. It is important to notice how the factor accurately mirror the recent crisis, in terms of timing. It indicates that the recession started around the end of 2007, beginning of 2008 and the restart of growth begins around the second half of 2009. This behavior seems very close to the one described by the GDP indicator of the StatFin dataset. It is a common problem to obtain monthly indicators of GDP and this factor seems to be a good proxy. The correlation between the GDP indicator and this factor is 0.85, which is a very high value. Factor 2 appears to replicate the first factor, even though the behavior of the two series is very different, almost opposite, in the beginning of the period of interest. Factor 2 can also be seen as an indicator of economic activity. Finally, the difference between factor 3 and the first two factors is more clearly defined. The last factor is much more volatile and more difficult to identify. Additional information can be obtained from the R^2 analysis, where I regress the micro factors on the variables of the StatFin dataset.

Variable	Factor1	Variable	Factor2	Variable	Factor3
GDP	0.74	OMX Telecomm.	0.13	HICP	0.25
IP	0.55	OMX Con.	0.12	CPI	0.16
Unemp. Job Seekers	0.47	GDP	0.12	Building Permits Tot.	0.15
Building Permits Res.	0.43			Reduced Time Work.	0.12
Purchase Price Inv.	0.33			Purchase Price Plants	0.10
3-Months Euribor	0.32				
Building Permits Tot.	0.313				
Reduced Time Work	0.312				
Eonia	0.30				
6-Months Euribor	0.29				

Table 10: R^2 of the Regression of StatFinDataset Variables on Estimated Factors.

From this table we can draw few intuitions. Firstly, factor 1 gathers most of the variation of the StatFin dataset. This can be seen by the very low R^2 of

the regression of the last two factors on the variables of the dataset, and from the very high R^2s related to factor 1. Another element we can draw from this table is the association of factor 3 and the price variables. It is worth pointing out that this is the first factor, including the ones of the other analysis, that can be easily interpreted as price indicator. Finally, the first factor can be easily associated with the overall economic activity, with very high R^2 for GDP, IP and unemployment related variables. Factor 2 is not identifiable and it does not carry much information about the variables of the dataset.

Finally, I present the plot of the cumulative eigenvalues for the micro dataset.

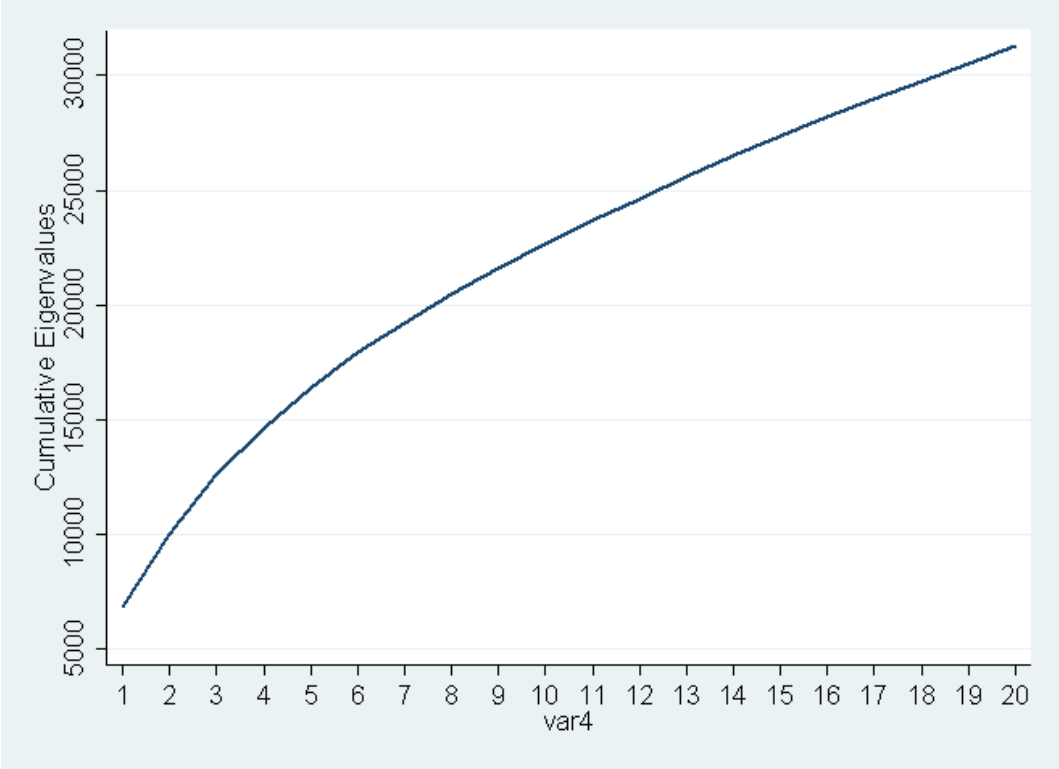

Figure 11: First 20 Eigenvalues, Cumulative

This plot presents analogies to the StatFin one. It seems that to explain the variance of the dataset, many factors are needed. This finding can be seen in contrast with the relative narrowness of the dataset. The micro dataset contains only turnovers, wages, prices and building permits, even though for many firms, product types and regions of Finland. This may indicate that cross-sectional variation can generate overall volatility of the data comparable to a dataset containing a very wide range of indicators.

8.3 Forecasting Results

The forecasting analysis follows the one of the StatFin dataset. The forecasting period, the variables to be forecasted and methods are equal to the ones of the first analysis. The difference stands in the fact that the factors used are estimated from the micro dataset. Again, the factors are extracted through a recursive method, where the micro dataset has been cut to adjust the time span to the one of the StatFin dataset (i.e. I cut the observations outside the period starting in September 2000 and ending in July 2010). Again, it is important to notice that I use the one year dynamic forecasts, instead of the classical 12-months ahead predictions. The results of the forecasts, in terms of relative $MSFE$, are reported in the next table.

Variable	Horizon	AR	VAR	Fac+AR	FAVAR	Fac(12)+AR
GDP	Dyn	1	0.74	0.49	1.16	0.49
	h=6	1	0.59	0.77	1.02	0.48
	h=3	1	0.62	1.36	1.18	0.52
	h=1	1	0.82	1.08	1.26	0.71
IP	Dyn	1	0.85	0.91	2.35	0.91
	h=6	1	0.80	0.77	2.003	0.92
	h=3	1	0.79	1.15	1.93	0.89
	h=1	1	0.92	1.08	1.614	0.89

Table 11: Relative $MSFE$ for Economic Activity Indicator Forecasts.

The performance of the forecasts is not as good as the one obtained in the first analysis. I believe that this is due to the fact that many indicators (financial variables, for example) are missing. Having said that, the performance of the factors-based models is significantly better for all the forecast horizons under exam, at least for GDP. The best model is the one using an autoregressive part (with two lags) and the 12th lag of the factors. The forecasts of the IP seem to be much more problematic, where only in one case the factors-based model performs better than the standard VAR. The forecasts plot is reported next. As before, I use the one year dynamic forecasts, using the factor plus AR model.

Figure 12: One Year Dynamic Forecasts for GDP and IP using Factors + AR Method

The plots show that the factor model based forecasts can predict the overall trend of the series, even though they underestimate the growth of GDP and IP. These results, in particular the overall ability of the factors models to replicate the trends of these series, were already suggested by the table of relative $MSFE$. The next step consists in examining the ability of these models to forecast the price series, CPI and HICP. Considering that factor 3 seems to explain a part of the variance of the price series, there are reasons to believe that the estimation of the factors-based model might generate better forecasts for the price series, at least compared to the StatFin based factors. The following table contains the relative $MSFE$ for the price series.

Variable	Horizon	AR	VAR	Fac+AR	FAVAR	Fac(12)+AR
CPI	Dyn	1	1.50	2.70	1.93	2.70
	h=6	1	1.45	1.86	1.91	2.59
	h=3	1	1.20	1.41	1.63	2.01
	h=1	1	1.142	1.33	1.43	1.37
HICP	Dyn	1	1.17	2.09	1.12	2.09
	h=6	1	1.18	1.42	1.19	1.98
	h=3	1	1.21	1.16	1.25	1.61
	h=1	1	1.27	1.13	1.30	1.26

Table 12: Relative $MSFE$ for Price Indicators Forecasts.

These results contradict my intuition. The models based on the factors perform worse in forecasting, where forecasting performance is evaluated by the relative $MSFE$. It is also worth noticing that factors-based forecasts are better for the HICP than for the CPI. The converse was true for the StatFin dataset, where the CPI was better forecasted than the HICP. Another element we can gather from this table is that the FAVAR model performs better for longer forecast horizons, compared to the other factors-based models. For the shorter horizons (3-steps and 1-step), the factors plus the autoregressive term model manages to create better forecasts. Having said that, none of the model manages to beat the autoregressive forecasts, which is the typical result of the literature. More insights might be gathered by looking at the plots of the forecasts.

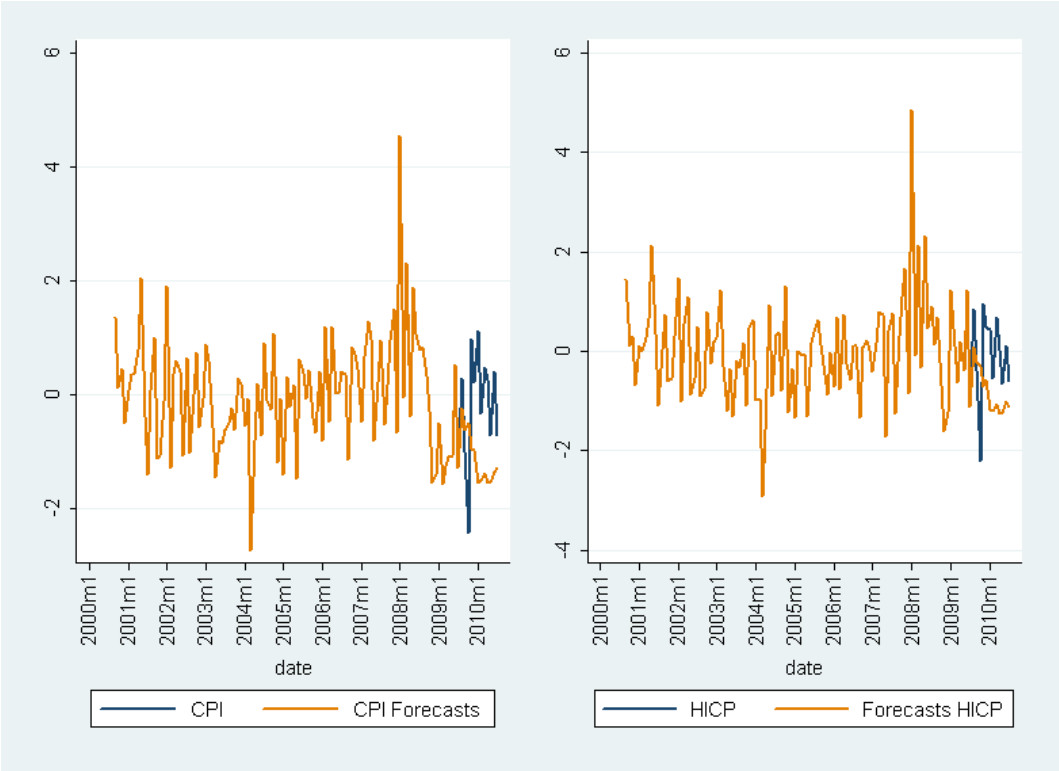

Figure 13: One Year Dynamic Forecasts for CPI and HICP using Factors + AR Method

These plots show that the models based on micro factors do not give good prediction results for price series. The forecasts do not manage to capture the overall volatility of the series, as they tend to underestimate the prices for all the forecast period. It is interesting to notice that the forecasts of the price series, based on the very volatile factors of the StatFin dataset do manage to capture the high variability of the series of interest. The micro factors, which show a lower degree of variability, generate forecasts that reflect this property.

This analysis confirms that the factors-based models create worse forecasts for price series than for economic activity series. The factors appear to be better at replicating the overall condition of the "real" side of the economy, while even for datasets containing many price indicators, the price level of the economy is not well represented by the factors. This characteristic seems to be in line with the findings pointed out by SW(2010): the gains in forecasting with factor models are greater for economic activity series. Still, it must be pointed out that the factors extracted from the StatFin dataset did a good job in forecasting CPI and HICP.

Before laying down the conclusions, it is interesting to have a look over the factors extracted from the three datasets of this study. I look at the correlation of the factors extracted from the datasets, to check how factors extracted from different datasets carry information. I separate the correlation tables for factor, thus I will include a table for the first factors, one for the second factors and finally one for the third factors. If the correlations of the factors are high in absolute value, it means that the SW method (2002b) can extract similar information for different dataset concerning the same economy. In theory, a robust method would create very similar factors, because datasets of a certain country should carry, to a certain degree, a common underlying movement. Of course, factors extracted from the BOF dataset have been cut, to ensure comparability with the factors derived from the other datasets.

Factor 1	StatFin	BOF	Micro
StatFin	1	0.61	0.58
BOF	0.61	1	0.49
Micro	0.58	0.49	1

Table 13: Correlation Between Factors 1 of Different Datasets.

The first factors extracted from the various dataset seem to be moderately correlated.The fact that the correlation is not extremely large, means that the datasets produce different factors. This is probably due to the presence of different variables and of different time span. The correlation between the StatFin factor 1 and the factor 1extracted from the BOF dataset is higher than the one estimated from the micro dataset. This could mean that the variable range is more important, in the estimation of the factors, compared to the time dimension. The same table, but for factor two is reported below.

Factor 2	StatFin	BOF	Micro
StatFin	1	-0.11	-0.29
BOF	-0.11	1	-0.08
Micro	-0.29	-0.08	1

Table 14: Correlation Between Factors 2 of Different Datasets.

The relation between factors 2 is completely different, with respect to the first factors. Now, the correlation are vary weak, if not moderately negative. Negative correlation is the sign that the factor can still be expressed as linear combination of one another, but here the negative correlation are not very high in absolute values. It seems that the estimation of the second factors is much more sensitive to the change in the datasets. Finally I report the table for the third factors.

Factor 3	StatFin	BOF	Micro
StatFin	1	-0.002	-0.06
BOF	-0.002	1	-0.09
Micro	-0.06	-0.09	1

Table 15: Correlation Between Factors 3 of Different Datasets.

Here, the correlations between the third factors is not far from being 0. The factors 3 carry very different information. From this analysis, it seems that the SW (2002b) method is more sensitive to the change in the data for factors corresponding to smaller eigenvalues, while the first factor seem to carry similar information, independently of the dataset employed in this thesis.
This analysis showed that the SW method (2002b) is not very robust to the datasets here used. Only the first factors are highly correlated, which means that they all capture a similar underlying characteristic of the various datasets. Looking at the plot of the graphs, and at R^2 analysis, the first factors seem to carry information about the overall economic activity of the economy. Factors 2 and 3, instead, carry very different information, depending on the dataset used.

9 Conclusions

The aim of this thesis is to provide a general overview at the dynamic factor model methods and their applicability to forecasting macroeconomic indicators in Finland. The forecasts were made following a two-steps procedure. In the first step I extracted the factors, using the static principal components method formulated by SW (2002b), and in the second step I used these factors in the forecasting equations. The factors-based models were compared to an AR model and to a VAR model. The evaluation of the forecasting performance was based on the relative $MSFE$, where the AR model was used as a benchmark.

The empirical analysis is based on three different datasets. The first one is a dataset containing macroeconomic indicators gathered from various Statistics Finland databases. The second one is formed of macroeconomic variables from Bank of Finland datasets. Finally, the last dataset is composed a number of micro (low aggregation) series. The results are dataset-dependent. The factor models functioned well in terms of forecasting for the first dataset, where both economic activity indicators and price indicators were well forecasted. The results are much less positive for the Bank of Finland dataset, where the factors-based models performed poorly. Finally, in the micro dataset, the models containing the factors produced good forecasts, at least for GDP and IP. In contrast, the price series, instead, were inaccurately predicted. However, the forecasts obtained in the StatFin analysis are more accurate.

This thesis is significant in two key respects. Firstly, it uses dynamic factor models to forecast Finnish macroeconomic variables, using large datasets. Secondly, and most importantly, it uses data with a very low level of aggregation. Indeed, to my knowledge, micro data have not been used in relation to factor models. The results obtained here give reasons to believe that further research following this line of enquiry might be worthwhile.

There are multiple research possibilities deriving from this analysis. One extension could involve analyzing the StatFin dataset across a longer time period. Finding longer time series for the micro dataset could also be crucial, in terms of establishing how reliable factors-based forecasts are over a longer time period (remember that I could not compute usual 12-step ahead forecasts for the StatFin and the micro

forecasting exercise). Related to the micro dataset, it could be interesting to use data at a more disaggregated level, for example at firm level. Furthermore, different line of research could derive from an analysis of the relation between the data and the extracted factors. By this I mean that it would be interesting to examine how the statistical properties of the dataset affect the estimation of the factors. This could be achieved through simulation methods. Micro factors could also have an application other than simply in forecasting. For example, a common problem for statistical agencies is that they have a number of missing values for firm-level data. One method that could be employed to fill these missing values, would be to use the fitted values of the regression of the various factors on the variable of interest. Finally, including factors in large structural models, such as used in central banks for forecasting purposes, could provide even better results. Including estimated factors in structural model (e.g. dynamic stochastic general equilibrium model, DSGE) is done in Boivin and Giannoni (2006). They show that exploiting the information contained in large datasets, using factor models, improves the estimation of the shocks effecting economic activity.

Dynamic factor models, and their relation to forecasting, have been a common topic of research in the last ten years and, as shown in this paper, there is a vast potential for future study, both on empirical applications and on theoretical results. This study contributed to the literature by showing that dynamic factor models have potential to create good forecasts for Finnish macroeconomic variables. Moreover, it has been shown that factors obtained by non-aggregated datasets are able to predict indicators of economic activity, even though the range of the variable is narrow. This suggests that this method is able to connect microeconomic and macroeconomic data, which has long been a source of interest for many economic theories and applications.

References

[1] ANGELINI, E., CAMBA-MÈNDEZ, G., GIANNONE, D., RÜNSTLER, G., REICHLIN, L.(2008) Short-Term Forecasts of Euro Area GDP Growth. ECB Working Paper 949.

[2] ARUOBA, S.B., DIEBOLD, F.X., SCOTTI, C.(2002) Real-Time Measurement of Business Conditions. *Journal of Business Economics Statistics* **27** 417-427.

[3] BAI, J., NG, S.(2002) Determining the number of factors in approximate factor models. *Econometrica* **70** 191-221

[4] BARTLETT, M.S (1938) Methods of estimating mental factors. *Nature***141**,609-611.

[5] BELVISO, F., MILANI, F.(2003) Structural Factor-Augmented VAR (SFAVAR) Preliminary.

[6] BERNANKE, B.S. and BOIVIN, J.(2003) Monetary policy in a data-rich environment. *Journal of Monetary Economics* **50** 525-546.

[7] BERNANKE, B.S, BOIVIN, J., ELIASZ, P.(2005) Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *Quarterly Journal of Economics* **120** 387-422.

[8] BOIVIN, J., GIANNONI, M.P.(2006) DSGE Models in a Data-Rich Environment. NBER Technical Working Paper No. 332

[9] BOIVIN, J. and NG, S. (2005) Understanding and Comparing Factor-Based Forecasts *International Journal of Central Banking* **1**, 117-151.

[10] BREITUNG, J., EICKMEIER, S.(2005) Dynamic factor models. Discussion Paper/ Deutsche Bundesbank **38** Frankfurt am Main: Germany.

[11] D'AGOSTINO, A. and GIANNONE, D. (2006) Comparing Alternative Predictors Based on Large-Panel Factor Models ECB Working Paper 680.

[12] DORSEY, R.E., MAYER, W.J.(1995) Genetic Algorithms for Estimation Problems with Multiple Optima, Nondifferentiability and Other Irregular Features, *Journal of Business and Economic Statistics*, 13(1), 53-66

[13] DOZ, C., GIANNONE, D., REICHLIN, L.(2006) A Quasi Maximum Likelihood Approach for Large Approximate Dynamic Factor Models, ECB Working Paper 674.

[14] EICKMEIER, S., ZIEGLER, C.(2008) How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach. *Journal of Forecasting*, **27(3)**, 237-265.

[15] FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L.(2000) The generalised factor model: identification and estimation. *The Review of Economic and Statistics* **82** 540-554.

[16] FORNI, M., LIPPI, M.(2001) The generalised factor model: representation theory. *Econometric Theory* **17**, 1113-1141

[17] FORNI, M., HALLIN, M., LIPPI, F., REICHLIN, L.(2005) The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting *Journal of the American Statistical Association* 100, 830-839.

[18] GEWEKE, J.(1977) The dynamic factor analysis of economic times series. Ch. 19 in AIGNER, D.J, and GOLDBERG, A.S. (eds.),*Latent variables in socio- economic models* Amsterdam: North Holland.

[19] GIANNONE, D., REICHLIN, L., SMALL, D.(2005) Nowcasting GDP and inflation: the real-time information content of macroeconomic dat releases. Finance and Economics Discussion Series 2005-42, Board of Governors of the Federal Reserve System (U.S)*Journal of Monetary Economics*, forthcoming.

[20] HAMILTON, J.D.(1994) Times Series Analysis. Princeton University Press. Ch. 13 372-408

[21] HENDRY, D.F.(1995) Dynamic Econometrics. Oxford University Press.

[22] KAPETANIOS, G., MARCELLINO, M.(2006) A Parametric Estimation Method For Dynamic Factor Models of Large Dimensions Discussion Paper/ Centre for Economic Policy Research **5620** London: UK.

[23] KAPETANIOS, G. (2007) Variable Selection in Regression Models using Non-Standard Optimisation of Information Criteria. *Computational Statistics and Data Analysis*, Forthcoming

[24] LAWLEY, D.N., MAXWELL, A.E(1962) Factor Analysis as a Statistical Method. *Journal of the Royal Statistical Society. Series D (The Statistician)*,Vol.12 ,No. **3** 209-229

[25] SARGENT, T.J, SIMS, C.A.(1977) Business cycle modelling without pretending to have too much a-priori economic theory. In SIMS, C.A. (ed). *New methods in business cycle research*Minneapolis: Federal Reserve Bank of Minneapolis.

[26] STOCK, J.H., WATSON, M.W.(2002a) Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics* **20** 147-162.

[27] STOCK, J.H., WATSON, M.W.(2002b) Forecasting Using Principal Components from a Large Number of Predictors. *Journal of the American Statistical Association* **97** 1167-1179.

[28] STOCK, J.H., WATSON, M.W.(2009) Forecasting in Dynamic Factor Models Subject to Structural Instability Ch.7. in Neil Shephard and Jennifer Castle (eds) *The Methodology and Practice of Econometrics: Festschrift in Honor of D.F. Hendry* Oxford; Oxford University Press.

[29] STOCK, J.H. , WATSON, M.W.(2010) Dynamic Factor Models. *Oxford Handbook of Economic Forecasting.*

[30] TIBISHIRANI, R.(1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, Issue 1 267-288.