Andrejs Bessonovs
University of Latvia, Latvia

THE ROLE OF NESTED DATA ON GDP FORECASTING ACCURACY USING FACTOR MODELS*

Abstract
The paper studies an impact of nested macroeconomic data on Latvian GDP forecasting accuracy within factor modelling framework. Nested data means disaggregated data or subcomponents of aggregated variables. The challenging issue regarding optimal number of macroeconomic variables to be used in factor models is pervasive since no criteria which states how many variables to employ and does disaggregated data improve factor model's forecasts. We employ Stock-Watson factor model in order to estimate factors and to make GDP projections two periods ahead. Several data incorporating schemes are tested whether it improves forecasting accuracy.
Results suggest that in the case of Latvia it's preferable to use the full database with all the subcomponents. Moreover results may improve if some preliminary data weighting scheme is applied.

Keywords: Factor model, forecasting, nested data, RMSE.
JEL: C22, C53, E37

1. Introduction

Seminal papers of Stock and Watson (1998, 2002a, 2002b), Forni and Reichlin (1998), Forni, Lippi, Hallin and Reichlin (2001a) put forward factor modeling framework as powerful tool to predict macroeconomic variables. Unlike the others univariate and multivariate models, factor models incorporate much macroeconomic data in the analysis. Stock and Watson (2002a) use 215 US macroeconomic variables covering the most economic sectors they may represent an economic activity and potential driving forces of an economy. Forni and Reichlin (1998) use 450 disaggregated series to understand aggregate dynamics.

Factor analysis is easy to implement by adding an additional data without any difficulty. The dataset may include as more information as more disaggregated time series are available for any additional specific sector of an economy. Since the former statement is logical to span the most sectors of the economy and to derive much variability from macroeconomic variables, whereas the latter is more uncertain and rises the question does the additional nested data brings more information to latent factors and hence enable to predict economic activity more accurate. Thus the goal of paper to study the problem of nested data and its contribution to forecasting procedures.

* Acknowledgement: this work has been supported by the European Social Fund within the project «Support for Doctoral Studies at University of Latvia». The opinion expressed in the paper is that of the author and do not necessarily reflect the opinion of the Bank of Latvia.

The paper of Boivin and Ng (2006) addresses the issue of the size and the composition of the data and its impact on factor estimates. They possess the question whether it is possible to obtain less useful factor estimates extracting them from larger datasets and argue that it is possible.

The paper of Caggiano et al. (2009) provides a comprehensive investigation on the factor modelling issues regarding number of factors, specification of the dynamics of the factors, combination of the factor-based forecasts and the choice of the dataset extracting the factors. Their empirical results point out that there are benefits of pre-screening of variables before extracting factors. For the raw of European countries pre-screening of the variables before estimating factors and then applying forecasting techniques improve forecasts substantially over the AR model benchmark. Caggiano et al. (2009) argue that the use about one fifth of original variables may yield the best results in terms of forecasts accuracy.

This paper is organizing as the following: in section 2 we describe a nature of data we use, any transformation and complexities capturing it in a model. Then the section 3 provides the model description and assumptions. Section 4 proceeds with obtained results and concludes the paper.

2. Data

We consider large dataset for Latvian economy with few additional time series of neighbor counties of Estonia and Lithuania. The data are collected on the main economic categories comprising business and consumer surveys of EU commission, industrial production, retail sales, consumer price indices, producer price indices, labour market, monetary sector, exchange rates, financial sector, foreign trade, fiscal sector and balance of payments (see Table 1). All the time series are with monthly frequency. Additional time series of Estonia and Lithuania are also included to keep dynamics of neighbor countries in common dataset making domestic factor estimates. These are real and nominal times series of industrial production, CPI components and confidence indicators of the main groups.

Table 1

Description of the databases and number of variables representing each sector

Full Database	Number of Variables
Confidence indicators	66
Industry	40
Retail trade	30
CPI	16
PPI	10
Labour market	2
Monetary sector	12

Exchange rates	4
Financial sector	8
Foreign trade	40
Fiscal sector	10
Balance of Payments	7
TOTAL	**245**

The most blocs of variables may contain data with high disaggregation degree. Consider total industry sector as in Table 2. It contains 3 main subcomponents: mining and quarrying, manufacturing and electricity, gas, steam and air conditioning supply. Moreover, manufacturing comprises manufacturing of food products, beverages and textiles etc. In turn, manufacturing of food products may contain even more disaggregated components. Thus the total industry represented by nests of some disaggregated parts.

Table 2
Representation of nested data for industrial production

Total Industry (BCD)
Mining and quarrying (B)
Manufacturing (C)
Manufacture of food products
Processing and preserving of meat and production of meat products (10.1)
Processing and preserving of fish, crustaceans and molluscs (10.2)
…
Manufacture of other food products (10.8)
Manufacture of beverages (11)
Manufacture of textiles (13)
…
Repair and installation of machinery and equipment (33)
Electricity, gas, steam and air conditioning supply (D)

Source: NACE rev.2.0

On the one hand, all those parts might be considered in a factor model all together. On the other hand, we can select any level of disaggregation and apply them further in the analysis. Besides, data choice may follow some selective manner based on any algorithm or criteria chosen by researcher.

The present study considers four schemes of databases' specifications. The first one (*Full*) is the full database comprising all 245 variables including all the aggregates and its subcomponents of all sub-levels. The second one (*Short*) is reduced-form database comprising mainly the first level aggregation. The nature of subcomponents time series is usually differs from those ones of

aggregates in the sense of volatility. Going deeper in disaggregate order we may find that those time series are more volatile because more specific sectors are more vulnerable to sector-specific shocks. Thus we leave the most aggregated variables in the second scheme and exclude subcomponents. Therefore judgmentally we reduce the full database to the sample of 54 variables. The next *Rule 1* scheme contains all the variables as in the *Full* database, but all the variables are weighted. Following Boivin and Ng (2006) the weighting scheme is defined as inverse diagonal elements of errors' variance-covariance matrix estimated from the factor model up to four factors. Intuitively every variable is weighted by the magnitude of the error variance to the total variance and basically it aims to account for heteroskedasticity in the errors. The last *Rule 2* scheme reduces the *Full* database to the smaller one by dropping variables with highly correlated error terms. In the case when variables are correlated with each other, then the variable with the highest R^2 is leaving. In turn, R^2 calculated regressing every variable on the four factors.

Time span of variables is from January 1996 to December 2010. All the variables are made stationary and normalized prior to factor estimation in order to neutralize differences in scale of variables (see Johnson and Wichern, 2007). The most of monthly series are subject to seasonal adjustment. Therefore all time series are seasonally adjusted by X-12-ARIMA method with specifications set by default, except interest rates and exchange rates, and those times series that already are available in seasonally adjusted form.

Data on Latvian gross domestic product (GDP) is collected on quarterly frequency. We compile real-time database in order to exclude methodology changes and GDP revisions effects on forecasting procedure (for details see Bessonovs, 2010).

Additionally the paper deals with the problem of missing values and ragged edge. Evidently, that all the monthly variables are supplied by statistical offices and respective officials with some delay or within individual schedule of publication as current month passes by. Therefore inevitably at any moment of time we observe ragged edge of data. The second problem arises as data not always is available for the desired period of time, especially at the beginning of the sample. The third, it might happen that few time series experience some breaks within the sample. These obstacles prevent us to implement factor estimation, because factor estimation techniques do not allow missing values. To tackle the problems above we apply expectation-maximization (EM) mechanism introduced in Stock and Watson (2002a) in order to achieve balanced panel of data. For additional information also see Bessonovs (2011).

3. Model

Similarly as in the paper of Stock and Watson (2002a) we employ the factor model. The general form of the model we set in the paper is the following:

$$y_{t+h|t} = \alpha + \beta_i' F_{t,N} + \sum_{j=1}^{p} \gamma_j y_{t-j+1} + e_t \quad (1)$$

Where $y_{t+h|t}$ is scalar forecasting value for *h* periods ahead, $F_{t,N}$ is a (*r*×*1*) vector of factor estimates using database of *N* series, y_{t-j+1} is y_t *j*-th lag variable, α and β_i coefficients.

Let the $X_t = (X_{1t}, \ldots, X_{Nt})'$ is the set of *N* variables at time *t=1,…,T*. Then the factors estimates, in turn, admit the following structure:

$$X_{it} = \lambda_i' F_t + u_{it} \quad (2)$$

Where X_{it} is *i-th* variable of database of *N* series (*i=1,…,N*), F_t is a (*r*×*1*) vector of factors, λ_i is (*r*×*1*) a vector of factor loadings for variable *i*, u_{it} is idiosyncratic error.

Concerning forecasting equation specification, note that for (1) we assume no any dynamics in factors and thus (1) is a static representation of factor model. In addition, to allow some dynamics of forecasting equation (1) we restrict *p=1*, i.e. there is one lag of dependent variable. Further (2) can be easily estimated by principal components and factors are the input for forecasting regression in (1).

As mentioned in section 2 the data frequency for monthly time series differs from GDP data and (1) cannot be estimated. To overcome that shortcoming we use (2) for monthly data, and then apply simple average function for monthly estimated factors to justify frequency basis.

4. Results and conclusions

In this section we compare the forecasting accuracy results. By means of root mean square error (RMSE) we measure magnitude of forecasting error as following:

$$RMSE = \sqrt{\frac{1}{T}\sum_{1}^{T}(\hat{y}_{t+h|t} - y_{t+h})^2}$$

where $\hat{y}_{t+h|t}$ is forecasting value at time *t* for *h* periods ahead, y_{t+h} is true value. Forecasting values and true values stand for year-on-year growth rates. The number *T* is set to be about 1/3 of available data sample size. Respectively 2/3 of actual sample is exploited for estimation and 1/3 for out-of-sample forecasting.

Results in Table 3 show RMSE for four data handling schemes with respect to AR(2) model results. Thus a number below 1 assumes factor model's better performance over AR(2) model. It also compares results among specified factor models with different number of factors.

Table 3
Factor models' RMSE results with respect to AR(2) model by the type of database

Out-of-sample forecasting period: 2005Q4-2010Q3				
1 period ahead	**SW1***	**SW2**	**SW3**	**SW4**
Full	0.72	0.70	0.72	0.75
Short	0.73	0.74	0.76	0.70
Rule 1	0.71	0.69	0.71	0.73
Rule 2	0.76	0.71	0.75	0.77
2 periods ahead	**SW1**	**SW2**	**SW3**	**SW4**
Full	0.78	0.77	0.78	0.79
Short	0.78	0.84	0.85	0.86
Rule 1	0.78	0.76	0.77	0.76
Rule 2	0.83	0.77	0.80	0.81

*Number denotes number of factors used in the model.

Results suggest that on average *Rule 1* scheme tends to outperform other data incorporation schemes for both time horizons ahead. We note also that both schemes Short and Rule 2 use reduced type databases in terms of number of variables and both show worse results comparing with complete databases' information.

Table 3 shows results comparing it with respect to AR(2) process. But we might be interested in to observe gains or losses in terms of percentage points. Again, comparison is worth to be against *Full* database, because this is the easiest way how to treat variables, just put all in the model. Therefore Table 4 gives the comparison of other schemes with respect to *Full* database by type of the model. Positive number states by how much certain scheme outperforms *Full* database in terms of average percentage points of year-on-year growth rates, respectively negative number states deterioration.

Table 4
Comparison of schemes' RMSE by type of the model

Out-of-sample forecasting period: 2005Q4-2010Q3				
	Improvement (+) / Deterioration (-)			
1 period ahead	**SW1***	**SW2**	**SW3**	**SW4**
Full	-	-	-	-
Short	-0.03	-0.16	-0.14	0.18
Rule 1	0.04	0.05	0.05	0.05
Rule 2	-0.16	-0.06	-0.09	-0.09
2 periods ahead	**SW1***	**SW2**	**SW3**	**SW4**
Full	-	-	-	-
Short	0.05	-0.53	-0.50	-0.48
Rule 1	0.01	0.07	0.07	0.23
Rule 2	-0.30	-0.04	-0.11	-0.11

*Number denotes number of factors used in the model.

According to the Table 4 forecasting 1 period ahead the *Rule 1* outperforms *Full* database on average by minor 0.05 percentage points and by 0.1 percentage points for 2 periods ahead. Other schemes perform worse and deteriorate results on average by 0.05-0.35 percentage points.

We have to admit that differences among the data schemes are rather small from practitioner's point of view. Nonetheless results suggest that the use of disaggregated components does not provide the evidence of huge efficiency loss or deterioration of the results due to disaggregated data. Moreover appropriately specifying the model efficiency gain is positive. Even more, the weighting data prior forecasting procedure might be advantageous.

Acknowledgement

ESF EIROPAS SOCIĀLAIS FONDS

LATVIJAS UNIVERSITĀTE ANNO 1919

IEGULDĪJUMS TAVĀ NĀKOTNĒ

This work has been supported by the European Social Fund within the project «Support for Doctoral Studies at University of Latvia».

Bibliography

1. Bessonovs, A. (2010) Measuring GDP forecasting accuracy using factor models: aggregated vs. disaggregated approach. Scientific Papers of University of Latvia, volume 758, pp. 22-33.
2. Bessonovs, A. (2011) GDP Modelling with Factor Model: an Impact of Nested Data on Forecasting Accuracy. MPRA Paper 30211, University Library of Munich, Germany.
3. Boivin, J., and Ng, S. (2006) Are more data always better for factor analysis? *Journal of Econometrics*, 132:169-194.
4. Caggiano, G., Kapetanios., G., and Labhard, V., (2009) Are more data always better for factor analysis? Results for the Euro Area, the six largest Euro Area countries and UK. ECB Working Paper, No. 1051.
5. Forni, M. and Reichlin, L., (1998) Let's Get Real: a Factor-Analytic Approach to Disaggregated Business Cycle Dynamics, *Review of Economic Studies* 65, 453-473.
6. Forni, M., Hallin, M., Lippi, M. and Reichlin, L., (2001a) Coincident and Leading Indicators for the Euro Area, *Economic Journal* 111, C82–85.
7. Johnson, R., A., Wichern D., W. (2007) Applied Multivariate Statistical Analysis. Sixth edition, Pearson Prentice Hall.
8. Stock, J.,H., Watson, M.W., (1998) Diffusion Indexes. NBER Working Paper No. 6702.
9. Stock, J. H., Watson, M. W., (2002a) Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business & Economic Statistics*, vol. 20, No. 2.
10. Stock, J. H., Watson, M. W., (2002b) Forecasting Using Principal Components from a Large Number of Predictors. *Journal of American Statistical Association*, Vol. 97, No. 460, pp. 1167-1179.