

Forecasting the Italian industrial production index using Generalized Dynamic Factor Model

Utilizzo del 'Generalized Dynamic Factor Model' per la previsione dell'indice della produzione industriale italiano

Fabio Bacchini, Anna Ciammola

Dir. Statistiche Congiunturali

Istituto Nazionale di Statistica

bacchini@istat.it, ciammola@istat.it

Monica Gentile

Dottorato in Economia e Management

Scuola Superiore Sant'Anna

gentile@sssup.it

Riassunto: Negli ultimi anni sono cresciuti considerevolmente gli studi e le applicazioni sui modelli fattoriali applicati a dataset estesi. In questo lavoro, utilizzando le serie degli indici della produzione industriale, studiamo il comportamento del 'Generalized dynamic factor model' proposto da Forni et al ([4]) sia nell'analisi del ciclo che in previsione. I risultati mostrano che le stime migliori dipendono dalla capacità di 'ripulire' il dataset dalle variabili scarsamente correlate con quella oggetto di studio.

Keywords: dynamic factor model, forecasting, industrial production.

1. Introduction

Krzanowski ([10]) identifies three main problems arising from the use of large data sets, that may require the selection or the removal of some variables: 1) the sample size (the number of observations) necessary to obtain reliable parameter estimates may be very high; 2) the number of parameters to be estimated may be too large yielding the problem computationally untreatable; 3) some of the variables may be irrelevant and/or may mask any important effects in the rest of the data.¹

The Generalized Dynamic Factor Model (henceforth GDFM), proposed by Forni *et al.* ([3]), represents a suitable tool in order to deal with large data sets and avoid problems such as 1) and 2). For the time being, however, it is not yet clear whether the removal of irrelevant variables could improve the final results. In this work, by using the data set of the Italian Industrial Production Indices (IPI), we deal with this last matter in two different contexts: the identification of leading and coincident indicators of IPI and the computation of forecasts. The innovative feature of the paper consists in our attempt to formalize the variable selection problem in the GDFM approach, looking at the forecast performance in terms of Root Mean Square Error (RMSE). Results are further investigated in order to estimate two different coincident indicator of the Italian industrial business cycle: one is based on the full data set and the other on the selected subset. This work is organized as follows: section 2 sketches the methodological foundations of the GDFM and describes the data set of IPI; section 3 deals with the variable selection by minimizing the RMSE of one-step ahead forecasts and discusses the results obtained in the estimation of a coincident and a leading indicator; section 4 concludes.

¹Indirizzo per corrispondenza: Fabio Bacchini, Istituto Nazionale di Statistica, Via Tuscolana 1776, 00173 Roma

¹On this point see also Boivin and Ng [3].

2. The GDFM and the data

Recently, the factor model techniques have received a remarkable interest in time series analysis and the availability of large data panels has facilitated their application in an increasing number of contexts ([5], [9]). Among these techniques a central role is given to the GDFM (Forni *et al.* [5] and [6]). Let $\mathbf{x}_{nt} = (x_{1t} \cdots x_{nt})'$ be a stationary n -dimensional vector process, with zero mean and finite second-order moments. In the GDFM approach each variable x_{it} can be expressed as:

$$x_{it} = \chi_{it} + \xi_{it} \quad (1)$$

where χ_{it} represents the common component and ξ_{it} the idiosyncratic term. The general idea of the GDFM is that under suitable condition, the process x_{it} has in common a small number of q aggregate indexes. Under the assumptions referred in the papers, Forni *et al.* provide two efficient estimators for the common component. The first one, proposed in [5], is a two-sided filter derived from a linear combination of past, present and future observations. The second one is an one-sided filter (built in two successive steps), introduced in [6] for forecast purposes. In fact it avoids the revision process of the estimates at the end of the sample as future observations are available.

The data set used in our application contains 85 time series referring to the IPI, for the period 1.1990-9.2003 in the base year 2000 ([8]). One series is relative to the general index, while the others refer to the groups (3-digits level of the ATECO nomenclature). In order to achieve the stationarity, the data have been transformed. Firstly we made the series free of outliers and deterministic effects (trading day, leap year and Easter effect) using TRAMO ([7]). Secondly we removed the seasonal component by means of dummy variables and applied the first difference. We did not use sophisticated procedures (as SEATS) to avoid the application of two-sided filters, which would imply large revisions in the seasonally adjusted series and therefore in the cyclical indicator². Finally, we posed zero mean and unit variance.

3. Application: forecast and coincident indicator for IPI

The transformed data set will be used to compute, by means of the GDFM estimation, *ex-post* forecasts of the general index for the span 10.2002 to 9.2003 and to estimate the coincident indicator. Moreover, we will try to understand whether the removal of redundant variables may improve the forecasting performance.

The computation of forecasts in the GDF framework, by means of the one-sided estimators, requires to set some parameters: the number q of common factors, the common factor lags s and the leads and lags l at which the covariance matrices has to be estimated. The parameters q and s identify the number of static factors used to compute the in-sample common components and/or the k -step-ahead forecasts (of the common components), where $k \leq |l|$. The forecasts of the idiosyncratic components, on the contrary, may be estimated by means of univariate ARIMA models. In our application, we have

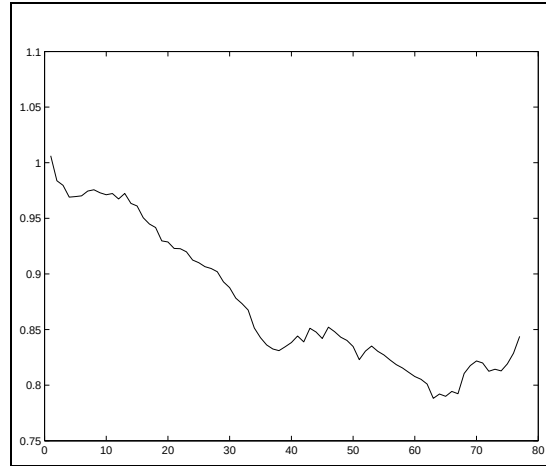
²This routine is the same proposed in Altissimo *et al.* [1]. However we investigated the transformed data in order to check the achievement of the stationarity, both running TRAMO in automatical manner and implementing Augmented Dickey Fuller (ADF) test. Results from TRAMO shown that 61 series exhibit a stationary seasonal component (in only one case with non stationarity component), while the ADF test always rejects the null hypothesis of non-stationarity.

identified $q = 1$, $s = 6$ and $l = 11$ minimizing the RMSE of the one-step-ahead forecasts (of both the common and the idiosyncratic components³) of the general index

$$RMSE = \sqrt{\frac{1}{12} \sum_{t=1}^{12} (IPI_{153+t} - \widehat{IPI}_{153+t|153+t-1})^2}.$$

where IPI_{153+t} is the observed value and $\widehat{IPI}_{153+t|153+t-1}$ is the one-step-ahead forecast. To reduce RMSE we try to formalized as a variable selection problem, whose objective function or selection criterion (to be minimized) is the RMSE. Clearly, the search of the best subset may be computationally prohibitive when the dimension n is large, as our application is. Among the various strategies that the literature proposes, we chose the backward elimination. This procedure works as follows: from the complete set of variables and the parameters q , s and l previously identified, each variable is omitted in turn and the RMSE evaluated; hence the subset of $(n - 1)$ variables that optimizes the criterion is found. Then, from this first subset, each variable is omitted in turn and the RMSE is computed; hence the subset of $(n - 2)$ variables is identified. This process is continued until the RMSE is falling. As shown in figure 1 as a local minimum when we delete 38 series and a global minimum when we delete 63 series.

Figure 1: RMSE values for different subset



Using the same dataset we built a coincident and leading indicator for IPI. The analysis runs in the following steps:

- identification of the number of q and cleaning of the variables;
- choice of the variables;
- classification of the series selected;
- building of coincident and leading indicator;
- detection of turning points.

Particularly to select variables for each serie we look at the ratio between the estimated variance of the common component and the variance of the series. We select only 33 series with the value of the ratio greater then 0.3 and then we re-estimate the common component for the selected series. Looking at the plot of the dynamic eigenvalues we

³The forecasts of the idiosyncratic component are estimated by TRAMO.

choose $q=1$.

As expected, 23 series as classified as coincident, 8 series as leading and 2 as lagging. Using the series weight in the industrial production survey, we calculate the coincident and leading indicator together with the turning points using Bry-Boschan algorithms.

For the coincident indicator we provide (Table 1) also a comparison with the turning points of the common component and with the turning points estimation by Altissimo et al. [2].

Table 1: Turning point estimation - 1.1993 - 1.2002

	P	T	P	T	P	T	P
ipi-comm	Jul-93	Aug-95	Nov-96	Jan-98	Apr-99	Aug-00	Jan-02
coin	May-93	Aug-95	Nov-96	Dec-97	Mar-99	Aug-00	Jan-02
Altissimo	Jul-93	Nov-95	Oct-96				

Results shows a good performance of the coincident indicators respect to the common component of the IPI. Also the leading indicators seems to anticipate the coincident indicator.

4. Conclusion

Empirical analysis suggests that the GDFM approach is useful both for business cycle analysis than for forecast. However results are sensible to the selection of the dataset.

References

- [1] F. Altissimo, A. Bassanetti, R. Cristadoro, M. Forni, M. Lippi, L. Reichlin, and G. Veronese. A real time coincident indicator of the euro area business cycle. *Temi di discussione, n.436 - Banca d'Italia*, 2001.
- [2] F. Altissimo, D.J. Marchetti, and G.P. Oneto. The italian business cycle: Coincident and leading indicators and some stilized facts. *Temi di discussione n. 377 - Banca d'Italia*, 2000.
- [3] J. Boivin and S. Ng. Are more data always better for factor analysis? *Manuscript*, 2003.
- [4] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized factor model: identification and estimation. *The Review of Economics and Statistics*, 82:540–554, 2000.
- [5] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. Coincident and leading indicators for the euro area. *The Economic Journal*, 111:62–85, 2001.
- [6] M. Forni, M. Hallin, M. Lippi, and L. Reichlin. The generalized dynamic factor model one-sided estimation and forecasting. *Manuscript*, 2003.
- [7] V. Gomez and A. Maravall. Programs tramo and seats, instruction for the user. *Working paper n. 9628 - Bank of Spain*, 1996.
- [8] Istat. Ricostruzione degli indici della produzione, del fatturato e degli ordinativi nell'industria anni 1990-2000. *Comunicato stampa*, 2003.
- [9] J.H. Stock and M.W. Watson. Diffusion indexes. *Manuscript*, 1999.
- [10] Krzanowski W.J. *Principles of Multivariate Analysis. A user's Perspective*. Oxford University Press (NY), 2000.