Journal of Multivariate Analysis 105 (2012) 348–367

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Estimation of high-dimensional linear factor models with grouped variables

Chris Heaton [a,*], Victor Solo [b]

[a] Department of Economics, Macquarie University, North Ryde, NSW 2109, Australia
[b] School of Electrical Engineering and Telecommunications, University of New South Wales, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 24 February 2011
Available online 18 October 2011

AMS subject classifications:
62H25
62H12

Keywords:
Factor analysis
Grouped variable approximate factor models
Instrumental variables
Time series

ABSTRACT

We introduce a generalization of the approximate factor model that divides the observable variables into groups, allows for arbitrarily strong cross-correlation between the disturbance terms of variables that belong to the same group, and for weak correlation between the disturbances of variables that belong to different groups. We call this model the Grouped Variable Approximate Factor Model. We establish identification, propose an estimation approach based on instrumental variable conditions that hold in the limit, and prove consistency in a dual limit framework. Monte Carlo simulations are used to investigate the performance of the estimator, and the techniques are applied to an analysis of industrial output in the US.

1. Introduction

Recent years have seen a renewed interest in linear factor models. The new literature is distinguished from the classical factor analysis of [29,32] and others by three main characteristics:

(a) Both the unobservable factors and the unobservable idiosyncratic disturbances are assumed to be serially correlated, weakly stationary time series processes that satisfy a weak correlation assumption in the time dimension.
(b) The idiosyncratic disturbances are permitted to be cross-sectionally correlated, subject to a weak correlation assumption in the cross-sectional dimension. That is, the model is an 'approximate factor model' in the terminology of [11].
(c) The analysis occurs in a framework in which $(N, T) \to (\infty, \infty)$, where N is the cross-sectional dimension of the observable vector, and T is the number of observations over time. Thus, the methods developed are intended for data sets that have a large number of observations on a large number of variables, and cases in which $N > T$ are tractable.

Stock and Watson [41] prove that the sample principal components corresponding to the largest k eigenvalues of the observable covariance matrix consistently estimate the normalized factor scores from a k-factor model, and that the Ordinary Least Squares technique may then be used to consistently estimate the factor loadings. Bai [3,5] prove asymptotic Gaussianity, and Bai and Ng [4] develop model selection criteria for consistently estimating the factor order k.

* Corresponding author.
E-mail addresses: chris.heaton@mq.edu.au (C. Heaton), v.solo@unsw.edu.au (V. Solo).

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.10.001

There now exist many applications of these techniques in economics. In particular, they have been applied to macroeconomic forecasting,[1] policy analysis,[2] index estimation,[3] and variance decomposition of aggregate variables.[4]

In this paper we present a generalization of the approximate factor model that may be more appropriate in some applications. In our model, each of the observable variables is a member of one of a fixed, finite number of groups. In different applications these groups might correspond to industry sectors, or geographical regions, or demographic characteristics; or they might simply be different categories of variable, e.g. real variables, nominal variables, financial variables, etc. There exists a set of factors which are common across all groups that we wish to estimate. We assume that a weak cross-sectional correlation condition is satisfied by the idiosyncratic error terms of variables that belong to different groups. However, in contrast to the approximate factor model, we allow for arbitrarily strong cross-correlation between the idiosyncratic errors of variables that belong to the same group. The model is completed by letting the number of variables in the groups grow large while the number of groups remains fixed. We call this model the Grouped Variable Approximate Factor (GVAF) model.

Our motivation for considering the GVAF model is based largely on a concern that the weak cross-sectional correlation condition that the approximate factor model places on the *entire* covariance matrix of the idiosyncratic errors may be too strong in some applications. As an illustration, consider an example where a factor model is used to estimate a variance decomposition of aggregate industrial output. Specifically, let

$$x_t = Bf_t + \varepsilon_t, \quad t = 1, \dots, T$$

where x_t is an $N \times 1$ vector of outputs in individual industries such that the elements of x_t sum to aggregate output. For simplicity, assume that f_t is a scalar common factor representing (say) the influence of aggregate demand on industry outputs. Since increases in aggregate demand are likely to affect all industries positively, assume $\exists c > 0$ such that $B_i > c$ for $i = 1, \dots, N$, where B_i is the ith element of the factor loading vector B. Without loss of generality, assume that $\Sigma_f = E(f_t^2) = 1$. The proportion of the variance of aggregate output that is due to variations in the common factor is

$$v = \frac{\omega' BB'\omega}{\omega' BB'\omega + \omega'\Psi\omega}$$

where $\Psi = E(\varepsilon_t\varepsilon_t')$ and ω is a vector of equal-valued weights that sum to 1. If highly aggregated industry variables are used, then N might be relatively small. For example, [33] use 13 industry output variables, Heaton and Oslington [25,26] use 9 and 10 industry unemployment variables. With finer levels of industry disaggregation, much larger values of N are possible. For example, [40] use employment data for 60 industry sectors, Foerster, Sarte and Watson [13] use industrial production in 117 industries, and [17] consider output and productivity in 450 industries.

Note that in the simple model presented above $\omega' BB'\omega = O(N^2)$. However $\omega'\Psi\omega \leqslant \sum_{i=1}^{N}\sum_{j=1}^{N}|\Psi_{ij}| = O(N)$ by the standard assumptions of the approximate factor model.[5] Consequently, if N is large, and the assumptions of the approximate factor model are taken seriously, estimation of v is redundant since its value is known to be approximately 1. This renders problematic the interpretation of variance decompositions constructed from large-dimensional factor models. If the finding is that almost all of the aggregate variance is due to the common factor, then this might be due to the modeling assumptions used to identify the common and idiosyncratic shocks, rather than being a reflection of the true composition of variance. On the other hand, if the finding is that the common factor does not account for most of the aggregate variance, then the appropriateness of the asymptotics of the principal components estimator of the approximate factor model might be questioned. We argue that, in this context, the GVAF model is an attractive generalization of the approximate factor model since it allows $\omega'\Psi\omega = O(N^2)$ and so does not constrain the value of v by assumption. More generally, in any application of large-dimensional factor analysis where the variables belong to a small number of identifiable groups and the 'large-N' assumption is satisfied by having a large number of variables in each group, the GVAF model may be more appropriate than the approximate factor model since it allows for arbitrarily strong error cross-correlation within each group. Monte Carlo simulations by Boivin and Ng [9] confirm that non-negligible correlation in the off-diagonal blocks of the error covariance matrix may seriously compromise the performance of the principal components estimator.

Since the entire error covariance matrix of the approximate factor model satisfies a weak cross-correlation assumption, any organization of the variables into groups will result in weak cross-correlation between the groups. Therefore, the approximate factor model is a special case of the GVAF model, and the theory and methods developed in this paper apply directly to it. Furthermore, any two groups in a GVAF model may be combined to produce a single large group that has errors which are weakly cross-correlated with those of other groups. Consequently, the GVAF model can accommodate situations where the number of groups grows with the number of variables, by simply combining the new groups with groups already

1 Breitung [10] and Stock[42] provide good surveys of the literature on forecasting macroeconomic variables with high-dimensional approximate factor models.

2 See, for example, [7,8].

3 See, for example, the monthly Chicago Fed National Activity Index (CFNAI) published by the Federal Reserve Bank of Chicago (http://www.chicagofed.org/economic_research_and_data/cfnai.cfm).

4 See, for example, [17,13].

5 See Assumption M1b of [41].