Journal of Banking & Finance 61 (2015) S177–S188

Contents lists available at ScienceDirect

Journal of Banking & Finance

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

Factor models for binary financial data

M. Fabricio Perez *, Andriy Shkilko, Konstantin Sokolov

The School of Business and Economics, Wilfrid Laurier University, 75 University Ave. W., Waterloo, Ontario N2L3C5, Canada

ARTICLE INFO

Article history:
Received 27 January 2015
Accepted 6 August 2015
Available online 13 August 2015

JEL classification:
G14
G30

Keywords:
Stock splits
Catering
Commonality
Factor analysis

ABSTRACT

Researchers are often interested in modeling binary decisions made by firms (e.g., the *yes* or *no* decisions to split the shares, initiate a dividend, or acquire another firm) as functions of economy-wide variables (common factors). Although factor models for continuous dependent variables are used widely, the toolkit of a financial researcher does not contain a generally accepted methodology that allows estimating factor models for binary dependent variables. In this paper, we study such a methodology. Using simulations, we identify data characteristics that allow for reliable estimates of factor parameters and conclude that the methodology is appropriate for the panel datasets of the type often used in finance. As an illustration, we use the methodology to address a currently debated issue of common factors in firms' decisions to split their shares.

1. Introduction

Linear factor models are often used in financial economics to model a panel of continuous observable variables as a function of a small set of unobservable (latent) factors. The estimation techniques for such models have been extensively studied. Recent advances in this area take advantage of the rich nature of financial data sets, where panel data often contain a large number of time series observations, *T*, and also a large number of cross-sectional observations, *N*. Bai (2003) shows that in such data-rich environments factor model parameters can be consistently estimated by principal component factors.

The advances in factor analysis have occurred in parallel with improvements in understanding of the links between economy-wide variables and firm behavior, including the links between (i) firms' cash holding decisions and credit market characteristics (Bates et al., 2009), (ii) firms' split decisions and market-wide investor preferences (Baker et al., 2009), (iii) firms' capital structure choices and various economic factors (Graham et al., 2014), among others. These studies construct proxies for common factors based on observable economic and financial data. For instance, Baker et al. (2009) use the relation between the market-to-book ratios of low-priced and high-priced firms to proxy for changes in investor preferences, while Graham et al. (2014) use the ratio of U.S. Federal debt to GDP to proxy for the supply of alternatives to corporate debt.

Using observable factors or constructed proxies is a widely accepted empirical exercise (e.g., Gagliardini et al., 2014). Nevertheless, in some cases there is no theoretical model that suggests an obvious observable factor or proxy. An alternative approach in such cases is factor analysis. The advantage of this alternative is that it does not impose common factors on the data, but rather lets the data reveal the factors if they exist.

Despite this advantage, there is a caveat in using factor models to examine some types of corporate decisions. Variables that capture these decisions are not always continuous; rather, they are discrete and often binary. For example, many issues of interest to finance researchers involve *yes* or *no* decisions made by firms; such as whether to carry out a stock split, initiate or increase a dividend, attempt an acquisition, refinance debt, etc. As we explain below, due to the binary nature of the variables of interest, estimation methods used for modelling continuous data do not automatically apply. Instead, discrete factor models are often estimated using MLE or weighted least squares techniques (Muthén, 1984; Jöreskog, 1990, 1994). These techniques are however not designed to deal with data-rich environments and cannot be applied when both dimensions of the data are relatively large. In this paper, we study a methodology that overcomes these issues.

We thank Alex Horenstein and the participants of the 1st Conference on Recent Developments in Financial Econometrics and Applications for comments and suggestions. Muhammad Khan provided excellent research assistance.

* Corresponding author.

http://dx.doi.org/10.1016/j.jbankfin.2015.08.012
0378-4266/

The methodology is simple and may be applied to data-rich environments (i.e., environments with large N and T) common in finance research. Specifically, we estimate factors using Bai's (2003) principal component factors (PCF) estimator, but based on the tetrachoric correlation matrix of the response variables instead of the Pearson correlation matrix commonly used for continuous data. The methodology can handle discrete response variables because tetrachoric correlations are designed for these types of variables and also exploits the benefits of data-rich environments needed for consistency of principal component factors.

After describing the methodology, we test its reliability using Monte Carlo simulations. Our simulation results show that the methodology provides very reliable estimators in data-rich environments. As expected, the reliability of the methodology is negatively affected when the factors are weak, and when the response variable is a rare event. This said, when the panel has a large number of observations as is often the case in financial data sets, these caveats are rather benign. Overall, our methodology works well when the number of cross-sectional units (firms) is over 1000, and the number of time periods is around 40 or more. For samples with the number of firms around 500, the number of time periods should be over 120.[1]

Many empirical research questions involve not only estimation of the common factors, but also estimation the true factor structure of the data (i.e., tests for the true number of factors). For example, in corporate finance an actively debated issue is the existence of common factors that drive firms' split decisions (Perez and Shkilko, 2015). For continuous data, several tests have been proposed to estimate the number of factors in data-rich environments (e.g., Ahn and Horenstein, 2013; Bai and Ng 2002). It is not *ex ante* clear if these tests accurately estimate the number of factors when applied to discrete data. Our simulations show that Ahn and Horenstein's (2013) tests perform well in discrete data environments if the researcher uses tetrachoric correlations, especially as the number of observations increases. In the meantime, using Pearson correlations leads to underestimation of the number of factors and to the incorrect conclusion that a factor structure does not exist.

To showcase our approach, we conclude with an empirical application. We use our methodology to shed new light on common factors in firms' split decisions. Baker et al. (2009) suggest that firms' decisions to split are driven by a common factor, a market-wide time-varying investor preference for low-priced stocks. Specifically, in some years investors may prefer buying stocks with low nominal prices. In such years, firms will split to appear cheaper and therefore more attractive to investors. Perez and Shkilko (2015) show that the variables in Baker et al. are highly persistent, likely leading to the spurious regression bias. When this bias is accounted for, firms' decisions do not appear to be driven by the preference for low-priced stocks. As such, the existence of a common factor that drives firms' split decisions is currently under debate. Our methodology is well-suited to revisit this issue and indicates that split decisions are entirely firm-specific. In other words, the data contain no evidence of a market-wide factor that systematically influences all firms' decisions to split their shares.

Our study complements several strands of a sizeable and continuously developing literature on estimation of factor models. Our contribution is to combine estimation methods designed for continuous variables in data-rich environments (Bai, 2003) with methods designed for discrete variables when one dimension of the data is small (Muthén, 1984; Jöreskog, 1990, 1994). We focus on factor models used for structural modeling. The idea behind structural factor models is that a set of observed variables is modeled as a combination of common and idiosyncratic components that are unobservable.[2] The objective is to estimate the true factor structure and the underlying factors and loadings. To the best of our knowledge, we are first to examine factor models with discrete data under the structural model setup in data-rich environments. This said, factor models are also commonly used for dimensionality reduction, where factor-based proxy variables are constructed to summarize the information from a large set of predictors.[3] Kolenikov and Angeles (2009) and Ng (2012) explore the effect of including discrete variables in factor analysis for dimensionality reduction and provide encouraging evidence on the use of PCF. Their objectives, methods and simulation design are different than ours, as we explain in the methodology section.

The remainder of the study is organized as follows. In Section 2, we describe the methodology. In Section 3, we report Monte Carlo simulations results. In Section 4, we apply the methodology to firms' decisions to split their shares. Section 5 concludes.

2. Methodology

2.1. Factor models for continuous response variables

We begin by examining the traditional factor model for continuous variables. This model was first studied by Spearman (1904). Let x_{it} be a continuous observable variable for subject i in period t. In finance applications, the subjects may include companies, managers, mutual funds, etc., and the time periods may be years, quarters, months, days or intraday intervals. The variables x_{it} are called *response variables* and may represent stock returns, financial leverage, managerial compensation, etc. A factor model assumes that the variation of x_{it} can be explained by a combination of (i) the common determinants across all individuals and (ii) the subject-specific determinants. Specifically, a linear factor model is defined as follows:

$$x_{it} = \alpha_i + \boldsymbol{\lambda_i} \boldsymbol{f_t} + \varepsilon_{it}, \tag{1}$$

where $\boldsymbol{f_t}$ is an $r \times 1$ vector of r unobserved common factors that represent variables that influence all subjects. The model assumes that these factors, although common, affect each subject differently, and these different effects are represented by $\boldsymbol{\lambda_i}$, a $1 \times r$ vector of factor loadings. In the meantime, ε_{it} summarizes the determinants of x_{it} that are subject-specific.

Several methods have been proposed to estimate the parameters of a factor model with an observable response variable x_{it}. The classic approach is to estimate the factor parameters using MLE and weighted least squares (WLS).[4] One notable characteristic of this approach is that consistent estimation of the parameters relies on the assumption that one of the two data dimensions (cross-sectional, N, or time-series, T) is fixed. As such, in practice these approaches are designed to work with samples where either N or T is small. Samples like these are often used in psychology, organizational behavior, marketing, and some other social sciences, but not as much in finance where large panel datasets are often available along both dimensions.

The latest developments in factor modeling relax the above-mentioned assumption, allowing for the sample sizes to be large

[1] The methodology is also reliable if the number of time periods is over 1000, and the number of cross-sectional units is around 40. Similarly, if the sample has 500 time observations, the number of firms should be over 120.

[2] The asset pricing literature models returns as a function of a small set of factors that are common to all firms and a set of firm-specific variables (1976 APT model). Similarly, firms' capital structure choices have been modeled as a function of common attributes (Titman and Wessels, 1988).

[3] Stock and Watson (2002) use common factors to summarize the information from 215 macroeconomic variables and use the resulting factor-based indexes for forecasting.

[4] Notable contributions to this area are by Anderson and Rubin (1956), Jöreskog (1970), Lawley and Maxwell (1971), and Browne (1984), among others.