COMSTA: 6185 Model 3Gsc pp. 1–19 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Revisiting useful approaches to data-rich macroeconomic forecasting☆

Jan J.J. Groen, George Kapetanios *

Federal Reserve Bank of New York, United States

Department of Economics, Queen Mary University of London, United Kingdom

ARTICLE INFO

Article history:
Received 27 May 2015
Received in revised form 17 November 2015
Accepted 23 November 2015
Available online xxxx

Keywords:
Macroeconomic forecasting
(weak) factor models
Principal components
Partial least squares
Bayesian ridge regression

ABSTRACT

The properties of a number of data-rich methods that are widely used in macroeconomic forecasting are analyzed. In particular, this analysis focuses on principal components (PC) and Bayesian regressions, as well as a lesser known alternative, partial least squares (PLS) regression. In the latter method, linear, orthogonal combinations of a large number of predictor variables are constructed such that the covariance between a target variable and these common components is maximized. Existing studies have focused on modeling the target variable as a function of a finite set of unobserved common factors that underlies a large set of predictor variables, but here it is assumed that this target variable depends directly on the whole set of predictor variables. Given this set up it is shown theoretically that under a variety of different unobserved factor structures, PLS and Bayesian regressions provide asymptotically the best fit for the target variable of interest. This includes the case of an asymptotically weak factor structure for the predictor variables, for which it is known that PC regression becomes inconsistent. Monte Carlo experiments confirm that PLS regression is close to Bayesian regression when the data has a factor structure. When the factor structure in the data becomes weak, PLS and Bayesian regressions outperform principal components. Finally, PLS, principal components, and Bayesian regressions are applied on a large panel of monthly U.S. macroeconomic data to forecast key variables across different subperiods, and PLS and Bayesian regressions usually have the best out-of-sample performances.

1. Introduction

It has been a standard assumption in theoretical macroeconomic modeling that agents are processing all the available quantities of information when forming their expectations for the future. Also, policymakers traditionally have looked at a vast array of indicator series in the run-up to major policy decisions, or in the words of Lars Svensson (Svensson, 2005) about what central bankers do in practice: '(l)arge amounts of data about the state of the economy and the rest of the world … are collected, processed, and analyzed before each major decision'. However, generally speaking it is either inefficient or downright impossible to incorporate a large number of variables in a single forecasting model and estimate it using standard econometric techniques. This prompted a new strand of research devoted to the theory and practice of alternative macroeconomic forecasting methods that utilize large data sets.

These alternative methods can be distinguished into two main categories. As, e.g., outlined in Hendry (1995), the methods of the first category involve inherently two steps: In the first step some form of variable selection is undertaken, including

☆ Supplementary material for this article is available online (see Appendix A).

* Corresponding author.
E-mail addresses: jan.groen@ny.frb.org (J.J.J. Groen), g.kapetanios@qmul.ac.uk (G. Kapetanios).

http://dx.doi.org/10.1016/j.csda.2015.11.014
0167-9473/© 2016 Published by Elsevier B.V.

automated model selection procedures as in Krolzig and Hendry (2001). The variables that are chosen are then used in a standard forecasting model. An alternative group of forecasting methods consists of estimation strategies that allow estimation of a single equation model that utilizes *all* the information in a large data set and not just an 'optimal' subset of the available predictor series. This is a diverse group of forecasting methods ranging from factor-based methods to Bayesian regression and forecast combination. We focus in this paper on the latter group of data-rich forecasting methods.

Within the group of data-rich forecasting techniques, factor methods have gained a prominent place. Building on Chamberlain and Rothschild (1983), Stock and Watson (2002a) and Bai (2003) show that under relatively weak assumptions regarding the behavior of the idiosyncratic components in a factor model, principal components can be used to identify the unobserved common factors in very large data sets. Stock and Watson (2002b) proved to be the starting point of a large empirical research output where, with mixed success, a limited number of principal components extracted from a large data set are used to forecast key macroeconomic variables.

Boivin and Ng (2006) make it clear, however, that if the forecasting power comes from a certain factor, this factor can be dominated by other factors in a large data set, as the principal components solely provide the best fit for the large data set and not for the target variable. This could explain why in some empirical applications principal components (PC) factor models are dominated by Bayesian regression and forecast combinations, as in both cases the information in a large data set is compressed such that this has explanatory power for the target variable. Under Bayesian regression one essentially estimates a multivariate regression consisting of *all* predictor variables, but with the regression coefficients shrunken to a value close to zero. Starting with Bates and Granger (1969), forecast combination involves the use of subsets of predictor variables in distinct forecasting models, which are then averaged to produce a final forecast. Note, however, that from an econometric perspective forecast combinations are *ad hoc* in nature.

Although less widely known, an alternative data-rich approach that can be used for macroeconomic forecasting using very large data sets is partial least squares (PLS) regression. We will show that PLS regression can do this irrespective of whether such a data set has a strong factor structure or not. PLS regression is implemented for large data sets through the construction of linear, orthogonal combinations of the predictor variables, which have maximize the covariance between the target forecast variable and predictor variables. Although similar in spirit to PC regression, the explicit consideration of the target forecast variable addresses a major existing criticism towards PC regression as a forecasting technique.

The main contribution of the paper rests on analyzing the properties of the various data-rich methods, in particular PC, PLS and Bayesian regression, under a more general setting for the target variable. In particular, most work to date has focused on modeling the target variable as a function of a finite set of unobserved factors. We, instead, assume that the target variable depends on the whole set of available predictor variables. As the number of these variables is assumed to tend to infinity this is a more difficult problem to handle. While some work (see, e.g., Stock and Watson, 2012) allows for such a setup, there are usually strict assumptions associated with this setup such as, for example, orthogonality of the regressors which both greatly simplifies the analysis and precludes interesting models such as factor models. We consider in detail the properties of PLS, PC and Bayesian regression for forecasting using both Monte Carlo analysis and an empirical application to gauge the potential of each of these data-rich approaches.

In the remainder of this paper we have the following structure: Section 2 discusses the asymptotic behavior of PC, PLS and Bayesian regression under different factor configurations: strong factors, strong factors underlying the predictor variables but only a few of these variables are relevant for the target variable, and weak factors. Section 3 report on an extensive Monte Carlo study that focuses on the out-of-sample properties of PLS, PC and Bayesian shrinkage regression. Section 4 presents an empirical application where PLS and the other data-rich forecasting methods are used on a large monthly US macroeconomic data set. Finally, Section 5 concludes.

2. Methods for data-rich macroeconomic forecasting

A useful framework for studying data-rich based modeling methods is provided by the following general forecasting equation

$$y_t = \alpha' x_t + \epsilon_t, \quad t = 1, \ldots, T, \tag{1}$$

where y_t is the target of the forecasting exercise, $x_t = (x_{1t} \cdots x_{Nt})'$ is a vector of dimension $N \times 1$ and thus $\alpha = (\alpha_1 \cdots \alpha_N)'$ is also $N \times 1$. The error term ϵ_t in (1) is throughout the paper assumed to be a stationary, finite variance martingale difference sequence. In this paper we will focus on the case that the number of indicator variables N is too large for α to be determined by standard methods such as ordinary least squares (OLS). The literature has proposed a number of ways how one can deal with this issue of large-dimensional data sets, of which we provide a selective review below.

Before proceeding, however, we need to stress that our assumed framework given by (1) is a significant deviation from, and, we argue, generalization of, the existing literature which analyzes the case where y_t does not depend on the large, observed data set, x_t, but a small, unobserved set of variables, referred to as factors.

We review methods that have been shown to be applicable for the data-rich case, starting with principal components (PC) regression in Section 2.1, partial least squares regression in Section 2.2 and Bayesian (shrinkage) regression in Section 2.3. In each subsection we present theoretical results and discussion on the properties of the methods under (1) while recapping the existing theory that relates to the usual factor setup utilized in the existing literature. Finally, we provide a short, qualitative comparison for the approaches in Section 2.4.

Please cite this article in press as: Groen, J.J.J., Kapetanios, G., Revisiting useful approaches to data-rich macroeconomic forecasting. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2015.11.014