Essays on Factor Models, Application to the Energy Markets

Ekaterina Ipatova

A thesis submitted for the degree of Doctor of Philosophy in Finance

Department of Finance, Cass Business School, City University London

Supervisors:

Doctor Lorenzo Trapani

Professor Giovanni Urga

April 9, 2014

This page is intentionally left blank.

Contents

List of Figures

List of Tables

Acknowledgements

At the outset, I would like to thank my principal supervisor Dr. Lorenzo Trapani. It has been an absolute honour to be his first PhD student. I appreciate his contributions in terms of time, intuition and guidance which helped transform my proposal into solid academic research. Without his support and patience my PhD would not be completed in time with the highest academic standards. Dr. Trapani's professionalism in the field has set an example which will always inspire me in my academic and professional career.

The advice and support of my second supervisor, Prof. Giovanni Urga has also been invaluable for which I am truly grateful.

My PhD would not be possible without the unwavering support of my mother and grandmother whose love and encouragement helped me progress. I thank them for raising me with a love for academic pursuits and engendering a curiosity in all things scientific.

And lastly, to Kaizad, the love of my life. Thanks ever so much for everything.

This page is intentionally left blank.

Declaration

I grant powers of discretion to the university Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgment.

This page is intentionally left blank.

Abstract

This thesis focuses on the development of the theoretical, methodological and empirical literature on factor models. We provide detailed descriptions of the techniques used to estimate factor models, as well a means to establish the number of factors and assumption of factor models. The opening chapters address research from the theoretical investigation, which is motivated by the fact that for the past fifty years theoretical econometricians were working towards relaxation of the assumptions and increasing the consistency of the estimators. We offer an alternative solution which engineers faster rates of convergence for the estimated parameters, and furthermore without imposing any additional assumptions.

The following chapter focusses on the problem of omitted observations in factor model datasets. Principle component analysis is only applicable to the balanced panel, therefore missing observations have to be filled. The modern literature predominantly focuses on the technique which can fill either missing observations at the beginning of the panel, or missing observations in the middle. Our technique offers a methodology which can help to fill missing observations irrespective of their place in the panel. Our technique is based on the factor model approach and uses factor model theory to develop the technique.

The closing chapter focuses on empirical application of the factor models. We attempt to assess forecasting ability of the factor models in comparison with non-factor augmented counterparts and the univariate model. We use a robust approach which has never been applied to factor models and the crude oil market. Ultimately we show that the factor model approach can significantly improve forecasting ability in the crude oil market.

This page is intentionally left blank.

This page is intentionally left blank.

1 Introduction

For more than 50 years the theory of factor models has retained a prominent presence in academic financial literature. The reason for the frequent application comes from the fact that factor models exploit the suggestion that a large number of series are driven by a limited number of common components. In other words, variations in the large number of market series can be adequately modelled by a small number of reference variables. A reduction in model variables helps to avoid the problem of reduced flexibility usually experienced with the regression based model. Additionally, measurement errors and local shocks can be estimated and excluded from the total variations. These advantages make the factor modelling methodology one of the most popular and powerful tools among researchers and practitioners.

In previous theoretical studies (e.g. Lawley and Maxwell (1971), Chamberlain and Rothschild (1983), Stock and Watson (2002), Bai (2003) and Bai (2004)) authors concentrated their efforts developing consistent estimators for the factors. During the literature review we identify a number of gaps in the existing theory. The purpose of our research is to fill these gaps by developing an inferential and predictability theory of factor models. In particular, in the first part of our research we propose a novel estimation methodology which aims to improve the robustness of estimated common factors, loadings and common components for the non-stationary panels of large time series (T) and cross sectional (n) dimensions.

Our research is motivated by the fact that the existing factor model literature does not differentiate the degree of consistency of the common factors estimated from levels as opposed to the first-differenced panels. Detailed examination of the optimisation of factor consistency provides an opportunity to make a contribution to the theoretical body of literature, reflecting the principle aim of the research. Specifically, we develop inferential and asymptotic theory for a novel methodology, showing that higher order terms converge to zero at a faster rate and (n,T) pass to infinity, suggesting that the proposed methodology yields better finite sample properties than direct estimation from first-differenced data. We describe methodology for Monte-Carlo simulation and empirical application that test developed theories. The details of the study along with rigorous proofs are presented in the second chapter of the thesis. The results of the study are applied to all further developments.

The third and forth chapters of the thesis concentrate on the methodological and empirical applications of factor models. Additionally, we present a possible practical application of the developed theory to the energy markets. Our literature review suggests that factor models theory has strong application and

would help solve current forecasting issues in the market. We consider this application as a possible addition to the main research focus, which is the development of factor models literature.

The third part of the thesis is motivated by the concurrent factor model literature failing to provide a unified methodology to overcome the problem of omitted observations in the large dimensional factor model panels. In the literature review we find number of studies that discuss possible means of filling missing observations (e.g. Baggi, Golinelli, Parigi (2004), Marcellino and Schumaher (2011), Foroni and Marcellino (2013)).Despite their attention, none of the studies where able to simultaneously fill missing observations at the end and in the middle of the panel including individual observations, and substantial missing blocks. We have to point out that the majority of the panels included omitted observations in the form of either individual missing observations, blocks of missing observations, mixed-frequency or "ragged edge" data.

Ideally, the methodology which should help to overcome a problem would enable a researcher to substitute all types of omitted observations present in the dataset, and take into account potential cross dependence of the variables due to the existing factor structure. Common techniques suggest extracting individual series and to substitute missing observations with cross-sectional independent variables. I have recognised a gap in the literature and attempt to provide a technique that is both simple to execute and one that can substitute any type of omitted observations in the factor based model. I employ factor models methodology to construct the EM- interpolation technique. Practical application accompanied by a rigorous proof allows me to distinguish and separate an accurate methodology from a parsimonious one. I demonstrate the validity of the technique by a number of Monte-Carlo simulation results, and empirical studies. I employ this technique to prepare the dataset for the final chapter which describes the empirical chapter of the thesis.

The final chapter provides a collection of the research ideas, contributing strongly to the academic literature as well as practical application. This is motivated by factor models never having been compared with alternative forecasting techniques in the robust framework. We use two factor models FA-VAR and FA-VECM to represent forecasting abilities of the factor model framework. The ARFIMA-GARCH models represent a univariate comparison model. We aim to establish the best forecasting model using the robust methodology described by Hansen (2011). The methodology uses the bootstraps technique to establish superior forecasting in the model and is able to mitigate the bias results of the simple loss function techniques. The loss functions, such as RMSE, MSE can determine the best result for the particular sequence of the data, however, these results may be drastically different in the future. Hansen's

et al. (2011) technique uses bootstraps which shuffle the data science and help mitigate this bias.

The third paper is an empirical study that determines the accuracyof the factor forecast in comparison to multivariate VAR and VECM models, as well as univariate models. We conduct the exercise on the WTI crude oil data for 1 month, 3 months, 6 months and 12 months to maturity future contract, representing future oil term structures. We use large dimensional panels of data that include the information about crude oil prises, supply, and demand determinants of crude oil prices and macroeconomic information. The original panel is unbalanced and we therefore use the findings of the third chapter to balance a panel used in the final part of the research.

The forth chapter reports on the results of the test of predictive ability of the factor models, multivariate models and univariate models. The results vary across the term structure, however, we can see that the factor approach demonstrates significantly higher levels of forecasting across the term structure. This we find to be the case for both short and long term forecasting. A detailed description of the work is presented in the forth chapter. The application of FA-VAR and FA-VECM models with information proxies can be extended to the other commodity markets; additionally, the EM-methodology can be applied to any factor base panel.

All three chapters work predominantly with the theory of factor models and, therefore, my thesis contributes the most to the factor models field. In view of the format of the final document, energy market research represents an empirical contribution in my findings. Overall the thesis contributes to the theoretical, methodological and empirical research on factor models. The first part of the research resulted in the development of an innovative technique that extends the research on factor models in the area of developing higher estimator consistency. The lliterature review indicated that over the past 50 years factor model research has moved towards more consistent estimators obtained from an unlimited dataset. This is a significant milestone in comparison to the original 1950s papers which imposed a number of restriction on the model, ensuring consistency. Moreover, datasets had to be finite along all dimensions. Modern theory is able to estimate common factors without imposing additional restrictions, also from unlimited datasets. Following this tradition of theoretical research which aims to improves estimation accuracy of the factors, our research amendments the theory and offers a way to improve estimation consistency even further without loss of generality. All the assumptions applied in the previous researcher of this topic is applicable and we have not imposed any additional constrains.

The third chapter contributes to the literature in so much as it provides the solution to the problem

of missing observations. Our contribution allows solving a problem of missing observations in the panel datasets, which are constructed for use in the factor model research. Based on the literature (see chapter 3) we distinguish between missing observations at the end of the panel, blocks of missing observations in the middle of the panel and individual missing observations. Current literature does not provide a methodology that can address all three types of missing variables. Our methodology attempts to fill this gap and offers a solution based on the application of the factor models theory and methodology. We have to impose additional assumptions to make sure this methodology is able to fill missing observations. Later research may concentrate on the means to relaxing the assumptions.

The final part of the thesis contributes to the empirical application of the factor models literature. The contribution of the final chapter focuses on the robust testing of the factor augmented multivariate models against univariate and multivariate counterparts. This has not been done before and we attempt to measure the validity of the factor model approach for the forecasting of the crude oil market. Additionally, the factor augmented vector error correction model has not been applied to the crude oil term structure; we attempted to contribute to the literature by evaluating the forecasting performance of the factor VECM model on crude oil market.

The remainder of the document is organized as follow: I begin the main body of the thesis by giving a historical summary of the main developments in the field of factor models. I describe the evolution of factor models, detailing estimation techniques and challenges during the process of finding the optimal number of factors. I also give a detailed overview of the existing literature on factor models, in addition to technical assumptions that are important for the consistent estimation of factor models parameters.

In the second chapter I describe the proposed novel methodology and demarcate difference between our model and existing ones. It develops a set of assumptions required for new estimators to be consistent. It describes the process of estimation and inferential theory, the development of asymptotic theory and shows the benefits of the new method. The closing parts of the chapter describe Monte-Carlo simulations and structural parameter evaluation used to demonstrate the gains of this novel methodology. The appendix presents proofs for deriving a limited distribution of factor loadings. Thereafter I present my concluding remarks and those areas of interest for further development.

The third chapter also provides a broad literature review on the history of the methodologies developed to overcome omitted variable bias, as well as a description of the modern solutions to the problem. I then move to describe a methodology of expected maximization approach. We use Monte-Carlo simulations

to test the results, along with empirical applications to the crude oil market and testing the structural parameters. At the end we provide a detailed summary and conclusion of the results. The fourth chapter provides a description of current developments on the crude oil market and details the motivations for researching the commodity markets. I describe models used in the "horse-race" and later present the results of the forecast evaluation for one and multistep forecasts. In the appendix I provide all previous theoretical findings in the form of the theorems, which are applied during the principle research developments.

This page is intentionally left blank.

2 Literature Review

2.1 Factor Models

For the past few decades, factor models have experienced increasing popularity in economic and finance related studies. This is primarily attributable to a growth in the availability of large datasets as well as advances in technology. The growing mass of information broadens the horizons for in-depth financial analysis, and modern technology helps to revolutionise the data processing techniques, which makes the analysis of large datasets both feasible and worth while. Growing application of large datasets presents new challenges for the traditional modelling practices, which experience the "curse of dimensionality" issues during the process of modelling large dimensional panels. Factor models offer an elegant mathematical solution to the problem by introducing a methodology, which reduces the dimensionality by searching for common patterns between variables (see,Forni et al.(2000)).

Many argue that pattern recognition, a bi-product of this technique, is a core reason for the popularity of factor models among practitioners, and in particular in the social sciences. Some indicators such as business confidence, are commonly treated as an easily quantifiable variable, despite them being qualitatively difficult to measure. The method by which to quantify such an hypothetical variable is to summarize the information from a large number of observable variables, by employing pattern recognition techniques in factor models. The latent factors of the models become a "measure index" of qualitative variables.

More rigorous fields, such as finance, employ pattern recognition to identify trends in the large time-series panels. In some cases it is possible to determine the latent variables with observable time-series. The model then bares a striking resemblance to the standard multiple regressions, for example traditional asset prising models, such as Arbitrage Pricing Model by Ross (1976), CAPM by Sharp (1964), or Gordon's triangle model for the inflation rate forecast (see Gordon (1988)). It is more intuitive to regress a panel of observable variables, such as multiple regressions. However, it is near impossible to find new observable variables (that have a strong correlation with the data) without a preliminary examination of main trends in the dataset. Factor models provide an ideal preliminary analysis of large datasets, which can subsequently lead to an interpretation of established trends. Even without direct interpretation, latent factors can demonstrate leading patterns of the data.

Pattern recognition is a crucial part of factor analysis, and it is the feature that is most commonly

confused with principal component analysis. For the benefit of current research, we would like to draw a clear distinction between two methods. According to Rencher (2002, p. 409): in principal components analysis (henceforth PCA), we define components as linear combinations of original variables. In factor analyses (henceforth FA), original variables are linear combinations of the factors. Additionally, PCA seeks to describe a large part of variation of the variables, but in factor analysis we account for covariance's between variables. Third, to apply PCA we do not require an initial set of assumptions. FA requires a number of assumptions, such as the covariance matrix is positive definite. Finally, PC produces unique components while in FA factors are subject to an arbitrary rotation. Factor analysis is preferred over PCA as we are able to find an interpretation of the latent variables. The interpretation of the factor is qualitative and aimed at providing best explanations to the set of common factors. Due to the fact that PCA does not allow the rotation, the interpretation of the factors is not easy.

2.2 Gaps in the literature

The theory of factor models covers a broad range of topics related to consistent pattern recognition techniques and further analysis of latent factors. During literature review we identify a number of gaps in the existing theory which we would like to address in this thesis. We were able to identify gaps in the related theoretical, methodological and empirical areas associated with factor models topics. We begin our investigation with an analysis of the theoretical literature developed for the factor models. These findings are especially relevant to the second chapter of the thesis. The third and forth chapters use the theoretical literature for consistent estimations of the common factors used in the methodological and empirical applications of the factor models.

In the second chapter we address the question of the development of the more consistent estimator for the common factor models. In previous studies (e.g. Lawley and Maxwell (1971), Chamberlain and Rothschild (1983), Stock and Watson (2002), Bai(2003), Bai(2004)) authors have concentrated their efforts on the development of the consistent estimators of the common factors. We provide a detailed historical overview of the theoretical development in the literature review below. Summarising the results, it is noticeable that theoretical research resulted in the development of the set of assumptions which secure the consistency of the estimators. The research gradually progressed towards the development of the more consistent estimators for the common factors (see Stock and Watson (2002), Bai(2003), Bai(2004)), with the most resent development on the topic establishing a set of assumptions which allowed to estimate

common factors and factor loadings which converge to the true factors of large dimensional panels. Our research continues a long line of the theoretical developments of the research which, aiming to improve the consistency and rates of convergence of the factor estimates. In this respect, our contribution ntends to improve the consistency in estimating factor models. rather than fill a hole in the existing literature.

The third chapter concentrates on the common problem in the large dimensional panel literature which relates to the problems of establishing a balanced panel of data when dealing with large dimensional panels. A detailed evaluation of the literature on factor models established that the present methodology of factor model estimation is only applicable to the balanced panels of data. However, while creating large dimensional panels the problem of missing observations is more prevalent than usual. This issue prevents direct estimation of the factor models and pre-requires filling of the missing observation. Over the course of the literature review we find number of studies that discuss possible options to fill missing observations (e.g. Baggi, Golinelli, Parigi (2004), Marcellino and Schumaher (2011), Foroni and Marcellino (2013)). However, none of the studies were able to simultaneously fill missing observations at the end of the panel and in the middle of the panel including individual observations and substantial missing blocks. Our study aims to fill this gap by developing an alternative approach which can help to solve the problem of differingt types of missing observations in the panel of data. This approach builds on general theoretical developments and a number of studies related specifically to the problem of omitted observations. General theoretical developments will be described below in the unified literature review, while specific literature on missing observations is addressed in detail in the second chapter.

The final chapter addresses the empirical challenge of forecasting using large dimensional factor models. In the past, a number of empirical papers have used large dimensional factor models to improve forecasting performance (e.g. Zagaglia(2010), Bernanke et al(2008)). However, the application of the large dimensional factor modes has never been tested in the robust framework. Our research fills this gap by comparing a performance of the factor vector autoregressive model and factor error correction model with a number of univariate models forecasts using Hansen (2011) bootstrap technique. Additionally, and to the best of our knowledge, the error correction model has never been applied to the energy markets, and it is for this reason that we are interested in evaluating the performance of the model using the robust framework. In a similar fashion to the previous chapters, the theoretical framework regarding the methodological estimation technique remains the same and its historical development is described in the generalised literature review below. A more specific review of the energy markets and factor models applied is given in the final chapter.

2.3 A critical summary of the main developments in the theory of factor models

Factor models were introduced by Lawley and Maxwell (1962) in the groundbreaking work on the statistical analysis of large dimensional datasets. The authors suggested compressing large dimensional data X by projecting the key trends of the dataset on the factor space F; while any variation from the main trends are classified as idiosyncratic noise. After performing the transformation, large dimensional data has a structure of equation 1:

$$X = \sum_{k=1}^{r} \Lambda F + E \tag{1}$$

Where F $(T \times k)$ is a matrix of common factors, Λ $(N \times k)$ is a matrix of factor loading; E $(N \times T)$ is a matrix of idiosyncratic errors; r is an optimal number of factors, which is substantially smaller than the total number of variables N. If r is large, then the model has not achieve a parsimonious description of the variables in a form of the function of a few underlying factors.

Basic assumptions are:

- The error terms are mutually uncorrelated, with $E(e_i) = 0$ and $E(e_i, e_i) = \sigma^2$. The assumption always holds when the dataset consists of stationary variables;
- Further we impose restrictions on the factors: $E(f_k) = 0$ and $E(f_k, f_t) = 1$;
- Additionally we assume independence of factors and idiosyncratic term: $E(f_k, e_{it}) = 0$.

The research on transformation of primary factor models reached a significant milestone when Chamberlain and Rothschild (1983) introduced the "approximate factor" model. The approximate factor relaxes the assumption of the primary model, such that it has an infinite number of column variables while time observations remain fixed. This improvement has resulted in a major reorganization of the theory of factor models. First, the model allowed for non-diagonal covariance matrix, which is not true for the primary model. Second, it demonstrates that PCA is equivalent to factor analysis when at least one dimension (N or T) goes to infinity.

The approximate factor model can be applied to broader sets of variables; however, it is still limited such that the covariance matrix $N \times N$ has to be known. In response to the problem Connor and

Korajczyk (1986, 1988, 1993) suggested estimating factor models using the covariance matrix $T \times T$ while N is larger than T. This amendment opens the research to the family of large dimensional models where both N and T tends to infinity with different speed. Such models of greater concern to the current study.

Stock and Watson (1999) were the first to describe static models with infinitely large N and T. The development of large dimensional models in linked with improvement in the quantity and quality of the data. The improvement in technical characteristics of high capacity computers helped to develop data collection and make processing easy and automated. Application of large dimensional panels allows for a more detailed analysis of current market trends resulting in a better forecasting ability.

The properties of large dimension factor models are different in comparison to the primary and approximate factor models. Therefore, new inferential and asymptotic theory of large dimensional static factor models have been developed to ensure consistent estimation in the model. The new theory established a more relaxed set of assumptions than was previously applied to finite models. Large dimensional factor models allow weak serial correlation in idiosyncratic terms, which are also generated by weak ARMA presses. In comparison primary models only allow for iid idiosyncratic terms. Homoskedasticity of idiosyncratic terms is significantly relaxed, as well as weak dependency between factors and idiosyncratic errors are permitted. This topic has been elaborated in the works by Bai (2003, 2004) where he derives rates of convergence and establishes consistency of estimated factors and loadings.

Current research is based on assumptions and model formulations from Bai (2003,2004). We distinguish between two types: stationary (large dimensional panel X constructed from I(0) variables) and non-stationary (large dimensional panel X constructed from I(1) variables). The model is given by equation 2:

$$X_{it} = \lambda_i F_t + e_{it} \tag{2}$$

$$F_t = \alpha F_{t-1} + u_t \tag{3}$$

Where X_{it} is variable in a matrix X $(T \times N)$ that contains a large dimensional set of variables; F_t is an observation in the matrix F $(k \times T)$ of common factors where k is optimal number of factors; λ_i is an observation in matrixΛ $(N \times k)$of factor loading; and e_{it} is idiosyncratic component; The common factor

is described by equation 3.

2.3.1 List of Assumptions

This section outlines assumptions applied to large dimensional factor models. The assumptions are taken from Bai (2003,2004) and they are specific to large dimensional factor models. For simplicity and consistency of our research with the rest of the literature, we retain the assumptions of Bai (2003,2004). Due to the fact that we use panels consisting of stationary or non-stationary series, than we provide two lists that separate the assumptions applied to either of the panel types.

We start from the assumption developed Bai (2004) p.140 for I(1) panels :

Assumption A (Common stochastic trends):

(i) $E||u_t||^{4+\delta} \leq M$ for some $\delta > 0$ and for all $t \leq T$;

(ii) As $T \longrightarrow \infty$, $T^{-2}\sum_{t=1}^{T} F_t^0 F_t^{0\prime} \underrightarrow{d} \int B_u B_u'$, where B_u is a vector of Brownian motion with covariance $\Omega_{uu} = \lim_{T\to\infty} 1/T \sum_{s=1}^{T}\sum_{t=1}^{T} E(u_t u_s')$;the $r \times r$ positive definite matrix and F_t^0 are true common factors

(iii) (iterated logarithms) $\liminf_{r\to\infty} \log\log(T) T^{-2}\sum_{t=1}^{T} F_t^0 F_t^{0\prime} = D$,where D is a non-random positive definite matrix ;

(iv) (initial value) $E||F_0^0||^4 \leq M$.

Assumption B (Heterogeneous factor loading) The loading λ_i is either deterministic such that $\left\|\lambda_i^0\right\| \leq M$ or it is stochastic such that $E\left\|\lambda_i^0\right\| \leq M$. In either case, $\Lambda^{0\prime}\Lambda^0/N \xrightarrow{p} \Sigma_\Lambda$ as $N \to \infty$ for $r \times r$ positive definite non-random matrix Σ_Λ.

Assumption C (Time and cross-section dependence and heteroskedasticity):

(i) $E(e_{it}) = 0$ $E|e_{it}|^8 \leq M$;

(ii) $E\left(\frac{e_s' e_t}{N}\right) = E(N^{-1}\sum_{i=1}^{N} e_{is}e_{it}) = \gamma_N(s,t)$, $|\gamma_N(s,s)| \leq M$ for all s, and $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\gamma_N(s,t)| \leq M$;

(iii) $E(e_{it}e_{js}) = \tau_{ij,s}$ *and* $(NT)^{-1}\sum_{i=1}^{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t=1}^{T}|\tau_{ij,ts}| \leq M$;

(iv) For every (t,s) $E\left|N^{-1/2}\sum_{i=1}^{N}[e_{is}e_{it}-E(e_{is}e_{it})]\right|^4 \leq M$.

Assumtion D $\{\lambda_i\}$, $\{u_t\}$, and $\{e_{it}\}$ are three groups of mutually independent stochastic variables.

Assumptions A to *D* outline a unique set of conditions that are applied to common and idiosyncratic sets of non-stationary factor models. To interpret assumptions we have to consider the norm of random matrix A to be denoted by $\|A\| = [tr(A'A)]^{1/2}$. Additionally, let F_t^0 and λ_i^0 be true common factors and true factor loadings respectively; M is a positive finite constant.

Assumptions A identify non-stationary common trends. *Assumption A(i,ii,iii)* defines distribution of the idiosyncratic factor u_t in the autoregressive process described in equation 3. *Assumptions A(ii,iii)* ensures convergence of idiosyncratic factors to a positive definite limiting matrix and thus rules out the possibility of co-integration between common factors. For more details on co-integrated trends see Bai (2004) who discusses identification and treatment of co-integrated trends in the non-stationary models. *Assumption A(iv)* ensures that size of fourth moment is bounded. *Assumption B* sets the properties of factor loading. In particular, it defines distribution up to the fourth moment and also that factor loading is always different from zero by setting positive the definite matrix of variance-covariance as N goes to infinity.

Assumptions C defines idiosyncratic component e_{it} in the factor model. *Assumption C(i)* relaxes normality condition of e_{it}. *Assumptions C (ii,iii)* allows for a limited time series and cross- sectional dependence between the error components, that lets a model to have approximate factor structure (see Chamberlain and Rothschild (1983). *Assumptions C(iv,v)* allow for Auto Regressive Conditional Heteroskedasticity (ARCH) in the error terms. However, the model performes better under homoskedasticity and the no-correlation condition. *Assumption D* rules out correlation between e_{it} and u_t.

Next, we outline assumptions for stationary factor models following Bai (2003) *Assumptions A-D*, p141.

Assumption E (Comon factors) $E\left\|F_t^0\right\| \leq M < \infty$ and $T^{-1}\Sigma_{t=1}^{T}F_t^0F_t^{0\prime} \xrightarrow{p} \Sigma_F$ for some $r \times r$ positive definite matrix Σ_F.

Assumption F (Factor loading) $\|\lambda_i\| \leq \bar{\lambda} < \infty$, and $\left\|\Lambda^{0\prime}\Lambda^0/N - \Sigma_\Lambda\right\| \to 0$ for some $r \times r$ positive definite matrix Σ_Λ.

Assumption G (Time and cross-section dependence and heteroskedasticity):

(i) $E(e_{it}) = 0, E\left|e_{it}\right|^8 \leq M$;

(ii) (Time-series dependence) $E\left(\frac{e_s' e_t}{N}\right) = E(N^{-1}\sum_{i=1}^{N} e_{is}e_{it}) = \gamma_N(s,t)$, $|\gamma_N(s,s)| \leq M$ for all s, and $T^{-1}\sum_{s=1}^{T}\sum_{t=1}^{T}|\gamma_N(s,t)| \leq M$;

(iii) (Cross-sectional dependence) $E(e_{it}e_{jt}) = \tau_{ij,t}$, *with* $|\tau_{ij,t}| \leq |\tau_{ij}|$ *for some* τ_{ij} *and for all t. In addition,*

$N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}|\tau_{ij}| \leq M$;

(iv) $E(e_{it}e_{js}) = \tau_{ij,s}$ *and* $(NT)^{-1}\sum_{i=1}^{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{t=1}^{T}|\tau_{ij,ts}| \leq M$;

(v) (Heteroskedasticity) For every (t,s) $E\left|N^{-1/2}\sum_{i=1}^{N}[e_{is}e_{it} - E(e_{is}e_{it})]\right|^4 \leq M$.

Assumption H (weak dependence between common factor and idiosyncratic errors):

$E\left(\frac{1}{N}\sum_{i=1}^{N}\left\|\frac{1}{\sqrt{T}}\sum_{t=1}^{T}F_t^0 e_{it}\right\|\right) \leq M$.

Assumptions E to *H* outline conditions applied to stationary factor models. *Assumption E* defines true common factors such that classical assumption of the strict factor model is relaxed (F^0 is i.i.d.) and some dynamics is allowed in the true common factor. It is however true that relations between true F^0 and X are still static. *Assumption F* ensures that factor loading has a unique contribution to the variance of a large dimensional panel. In our research factor loadings are uniformly distributed. *Assumption G* are equivalent to *Assumption C* which describes conditions for the idiosyncratic factor. *Assumption H* allows weak dependence between factors and error components.

2.4 Critical evaluation of the methodology of factor model estimation

Factor model estimation can be described as a two step procedure. First we derive the factor model parameters and determine their optimal number. Second we estimate factor scores by linear regression. Modern literature recognises two widely applicable methods for the estimation of factor model parameters. Maximum Likelihood (ML) technique was the original solution and later Chamberlain and Rothschild (1983) suggested Principal Component Method (PC) as an alternative. The choice between two techniques

depends on the individual characteristics of the data. According to Rencher (2002) application of ML is largely limited by the fact that it is only available for relatively normally distributed and finite data. However, given that these characteristics are satisfied ML is preferred. This is partly due to the fact that ML is an extremely computationally efficient method for parameter estimation, and it is easy to obtain additional statistics that help to assess the statistical significance, goodness of fit, confidence intervals for factors and evaluate correlation between them.

Non-normality and high dimensionality of the data presents a problem for the traditional ML estimators. More precisely, ML estimator fails to converge when the number of parameters for assessment tend towards infinity. Non-normality of the data leads to biased and inefficient estimations. To solve these problems, Principal Component Method was introduced. Due to the nature of the present research we choose to use Principal Component Method for estimation of model parameters. First it overcomes the limitation of ML and can be applied to the non-normal data, which provides data restriction in the research. Second, the analysis is set to work with large dimensional panels and ML estimators can fail to converge when the number of estimated parameters is going to infinity. Third is that modern literature on large dimensional factor analysis suggests PC for factor estimation. Finally, Principal Component Method is most commonly applied to static panels, which are the subject of this research. Due to the importance of the PCM we provide a complete methodology of the approach in the following chapter. It is true that Principal Component Method has a number of variations such as Principal Axis Method (PAM) and Iterated PAM. We therefore feel that it is essential to define the differences between the various approaches and justify the reasons for choosing the PC. All the analysis on the topic is provided in the following chapter.

Although Maximum Likelihood has no direct application in the current research we provide a summary discussion about this methodology, in order to assess the various techniques available for factor analysis in finance. We also acknowledge the alternative and varied techniques additionally available to carry our factor analysis (see Alpha Factoring, Image factoring, Regression factoring using OLS/ GLS, Bayesian regression constructed by De Mol et al. (2006). However these methods are rare in economics and, therefore, we refrain from carrying out a detailed discussion on the same.

It is essential to point out that implementation of Principal Component Analysis demands further restrictions in the form of assumptions for available data. The list of assumptions has evolved over the past few decades and current research considers the latest developments. We present the summary discussion regarding individual assumptions.

2.4.1 Common Factor Estimation

Principal Component Method and Related Approaches *Principal Component Method* (PC) was introduced as a technique available primarily to large dimensional factor models, where the matrix of errors is not normal. PC approximates common factors (F) by applying spectrum decomposition to the covariance matrix S:

$$\begin{aligned} S &= \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' / n \\ S &= C\Theta C' \\ \hat{S} &\doteq \left(C\Theta^{1/2}\right)\left(C\Theta^{1/2}\right)' = \hat{F}\hat{F}' + \hat{\Psi} \end{aligned} \tag{4}$$

We start by estimating covariance matrix of the original data using equation 4. We attempt to factor S on normalised eigenvectors C and diagonal matrix of eigenvalues Θ. To do that we apply spectrum decomposition and extract eigenvalues that are roots of equation $|S - \Theta I| = 0$, where I is an identity matrix and S is a positive semi-definite matrix. At the later stage we attempt to find eigenvectors C by solving system of linear equations $|S - \Theta I| C = 0$, where Θ is know from the previous stage. Eigenvectors C are normalised by dividing each non normalised eigenvector by its length. Normalization shifts the coordinate axes to a new coordinate system, thus common factors become a convenient set of coordinates. Finally, we neglect matrix Ψ and attempt to determine common factors using third equation.

Matrix Ψ assumed to be diagonal and equal to $diag(\psi_1, \psi_2, ...\psi_n) = I - diag(f_{11}^2, f_{22}^2, .., f_{nn}^2)$, all nondiagonal elements of Ψ assumed to be equal to zero. The error matrix $\Phi = S - (\hat{\Lambda}\hat{\Lambda}' + \hat{\Psi})$, s.t. diagonal elements Φ are equal to zero, but off-diagonal elements are non-zero. The proportion of total variance of S due to the fact that each factor is estimated by dividing each eigenvalue on trace of covariance matrix $\vartheta_i / tr(S)$.

Principal Component Method has a number of variations that can be found in the modern literature. Principal Axis Method (PAM) is possibly the most common variation of PC and it also can be applied for the smaller panels. To apply PAM we start from the common expression $S = \hat{F}\hat{F}' + \hat{\Psi}$, however unlike the PC we do not neglect matrix Ψ, but attempt to approximate factor loadings using matrix $S - \Psi$ instead of S. Principal Axis Method can be easily transformed to the Iterated PAM. The first stage will be to compute factors using standard PC, and then use new factors to approximate matrix $S - \Psi$. The

process then goes back to the first stage and continues until the convergence.

The difference in the methodologies is marginal and as a result all these methods lead to the similar results providing that (i) correlations are fairly large, with a resulting small value of k; and (ii) number of variables n is large. According to the nature of the research it can be observed that application of any of the methods would lead to similar results. However, Principal Component Method is computationally easier therefore it is our main methodology for factor extraction.

Maximum Likelihood Factor analysis was originally performed using Maximum Likelihood (ML) technique (see for example works by Sargent and Sims (1977), Stock and Watson (1989)). Principle Component method partially evolved from ML and it bares certain similarities to the approach that is still widely recognised as a valid technique for factor approximation. In this chapter we provide a short description of ML because it (i) carries historical importance for factor analysis and (ii) we outline ML procedure to be able to compare and describe the benefits of PC approach as a technique that was developed to overcome the difficulty posed by ML. The procedure described below follows Lawley and Maxwell (1962) original desings, and the more current work discussed by Bartholomew (2011).

$$\begin{aligned} S &= \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' / n \\ \hat{S} &= \hat{F}\hat{F}' + \hat{\Psi} \\ L &= -\frac{1}{2}\ln|\hat{S}| - \frac{1}{2} n \sum_{i,j} a_{ij} c^{ij} \end{aligned}$$

We launch the ML technique by estimating covariance matrix S for panel X_{it}. It is obvious that variance-covariance matrix can be approximated using common factor F as well as estimated directly from the panel using the first equation. We use this idea to build likelihood function L, that we maximize by setting partial derivatives (with respect to f_{ir} and Ψ_i) equal to zero. The equation is estimated iteratively up to convergence of the covariance matrix.

To be able to perform the procedure we have to impose number of conditions that tend to restrict application of ML. First and most importantly, data has to be normally distributed $X \sim N_p(\mu, \Sigma)$. Second, ML fails to converge given that covariance matrix is too large (for example matrix S $T \times T$ when $T \to \infty$) see Lawley and Maxwell (1962).Given that n is finite, it is possible to apply ML for

consistent estimation of factor loadings but not common factors. Common factors are then extracted using regression $\hat{F} = XS^{-1}\hat{\Lambda}$. Next, we have to include conditions such that: (i) the matrix of common factors F is independent from the matrix of errors e_{it};(ii) the matrix of errors should be independent and identically distributed across time and independently across i; (iii) finally the matrix $\Omega(e_t e_t')$ is diagonal. These assumptions are very strict for finance and economics data and therefore the Principle component approach was developed to overcome such restrictions.

2.4.2 Optimal Number of Factors

The problem of factor extraction is related to the question of the optimal number of factors (k), which can be estimated using a number of techniques. In this section we give an overview of the most commonly applied techniques, including the literature review on the topic and the reasoning behind favouring specific methods.

We start from Bai and Ng (2002) approach for estimation of optimal number of factors, which is based on information criteria. The information criteria optimizes the number of maximum possible common factors kmax in the model as a trade-off between accuracy of fit and over fitting, where $k\max < \min\{n, T\}$ and $k < k\max$. Information criteria is estimated using the following formulas $PC(k) = V(k) + k * g(n, T)$, where $V(k)$ is the minimized squared residuals, k is number of factors and $g(N, T)$ is a penalty function. For the factors estimated from the first-difference data, information criteria should be estimated with the equation: $PC(k) = V(k) + k\sigma^2 \left((N+T)/NT\right) \ln\left(NT/(N+T)\right)$, where $\sigma^2 = V(k\max)$. The number of factors for data in levels is calculated using the following information criteria: $PC(k) = V(k) + k\sigma^2\alpha_T \left((N+T)/NT\right) \ln\left(NT/(N+T)\right)$, where $\alpha_T = T/\left[4\ln\ln(T)\right]$ by the law of iterated logarithms. Information criteria computes the number of factors consistently only for large dimensional datasets.

There exist a number of alternative approaches. The classical approach for factor number selection is *the variance based approach*. To perform this we first have to select the optimal amount of variation that has to be explained (usually between 80%-90%). Total variation is estimated using $tr(\Sigma)$, where the amount of variation of each loading is equal to $\vartheta_i/tr(\Sigma)$, where ϑ_i is eigenvalue of each factor loading. We choose optimal number of loading k, so that the sum of explained variation constitutes a relatively large portion of total variation. It can also be estimated as a sum of squares of all elements of $\hat{\Lambda}$ to $tr(\Sigma)$.

The next method applies *Kaiser criterion (*see Bandalos and Boehm-Kaufman (2009)*)*, which omits all factor loadings with eigenvalues smaller than average, that is $\sum_{n=1}^{N} \vartheta_i/N$. While applying Kaiser criterion to PAM we omit all negative eigenvalues, which tend to reduce too many factors.

Thee *Cattell screen test* is based on the plot of each loading on the X axis and its corresponding eigenvalue of Σ on Y axis, in which all values are sorted in an ascending order. As we move to the right we observe the eigenvalue to drop and form a curve. We choose all eigenvalues (with their corresponding factors) which are located before a noticeable sharp drop in values of the curve. This test is very popular among practitioners, and the selection is based on a visual assessment of the plot.

Forni et al. (2000) proposed the use of an heuristic examination of factor loadings against variables n. The minimal percentage of variance has to be explained by each prespecified factor. Therefore, the number of first q eigenvalues converge to the true one, while $N - q$ remain bounded by original pre-spesification and can be neglected. This approach is highly sensitive to the prior specification and can therefore yield biased results.

2.4.3 Interpretation of common factors: Rotation

The theory of factor models recognises estimation of unobservable factors (see Stock and Watson (2002)). However, practically applied, it is common to seek interpretation of those factors. To do this we intend to group original variables in to clusters formed on the basis of the largest values for each factor. For example, if the first factor is a vector (.927, -.037, .980 , .916, .194) then the first, third and forth variables represent a cluster. By analysing the first, third and forth variables from the original panel X, we can find an interpretation by looking for commonality between them. However, the values of factors do not always present clear clustering and in this situation we have to apply a rotation technique. The rotation aims to separate factors which contradict each other and make the model more interpretable. In addition, we aim to reduce the number of negative factors which are hard to explain; finally, we reduce as many factors to zero in order to reduce parameters of the model.

The orthogonal rotation is the most commonly applied. Graphical interpretation of the rotation consists in moving the factor (axis) closer to the cluster(s) of factor values. Rotated factors form a similar covariance matrix, and can be easily interpreted.

$$S = \hat{F}^* \hat{F}^{*\prime} + \hat{\Psi} = \hat{F} T T' \hat{F}' + \hat{\Psi} = \hat{F}\hat{F}' + \hat{\Psi}$$

To estimate $\hat{F}^* = \hat{F}T$ we have to multiply the matrix of factors by the rotation matrix, that is a matrix of Sin and Cos functions that determines the rotation angle. Graphical rotation is only possible for simple systems with two factors. In more complicated situations we apply varimax rotation that looks for rotated factors maximizing the variance of squared factors in each column $\hat{F}^*$. As a result we obtain a matrix of factors that can be clearly grouped into clusters.

2.4.4 Factor Scores Estimation

To complete the factor models it is crucial to estimate factor scores. The procedure was developed by Bartlett(1938) and it builds on the idea of minimisation of the sum of squared standardised residuals $\sum_i e^2/\psi_i$, where ψ_i are from equation $\hat{F}\hat{F}' + \hat{\Psi}$. The sum can be rewritten as follow:

$\sum_i e^2/\psi_i = \sum_{i=1}^{N} (x_i - \sum \lambda_{ik} f_k)^2/\psi_i$

We minimize the above equation with respect to F to arrive to simple regeression $\Lambda = X \times F/T$. Following Stock and Watson (2002a) we distinguish between two cases of "short panels" where $T < N$ and "long panels" where $T > N$. For short panels we apply previously described computational technique, i.e. we estimate common factors by PC and later estimate factor loadings using regression $\Lambda = X \times F/T$. For "long panels" where $T > N$ we construct the loading by applying PC to $N \times N$ matrix $X'X$ and the common factor is computed using regression $F = (X \times \Lambda)/N$.

This page is intentionally left blank.

This page is intentionally left blank.

3 First-Differenced Inference for Panel Factor Series

Ekaterina Ipatova Lorenzo Trapani

Cass Business School, City University London

April 9, 2014

Abstract

The existing inferential theory for non-stationary panel factor models is extended by proposing a novel estimation methodology for common factors, loadings, and common components, in the context of large time series (T) and cross sectional (n) dimensions. The method proposes to extract the non-stationary common factors by applying Principal Components (PC) to data in stages, and then uses their first differences. First order asymptotics of the estimated loadings and common components are found to be the same when the stationary factors are directly estimated using first-differenced data. Conversely, higher order terms are shown to converge to zero at a faster rate (n,T) and pass to infinity, thereby suggesting that the proposed methodology yields better finite sample properties than direct estimation from first-differenced data. The theoretical findings are investigated through the comprehensive Monte Carlo exercise, showing that even in the case of small N and T, the asymptotic results form a very good approximation of the finite sample properties of the stimated loadings and common components..

3.1 Introduction

Previous studies on factor models theory concentrated on identifying those conditions that allow the estimated factors to be treated as 'known' and 'true', that is, when the estimation error is negligible; that is when the estimation error is negligible. Of the early theoris, Lawley and Maxwell (1971) specified a list of strong assumptions that may be applied to a limited data sample, aiming to ensure convergence between the estimated factor and the true theoretical trend..By partially relaxing the assumptions, as Chamberlein and Rothschild (1983) did with their theory of large dimensional factor models, one is able to ensure convergence between the estimated parameter and the true trend for much larger data samples. The complete theory of large dimensional factor models was established twenty years later when Stock and Watson (2002), Bai(2003), Bai(2004) identified theoretical properties of the estimators of the large multidimensional factor model, in addition to the list of assumptions which guaranteed convergence between the parameters and the true factors.

The development of the theory of large dimensional factor models coincided with growing technological progress, and had a huge impact on empirical research findings. Newly available large data sets demand an analytical tool which can access and extract the information core locked in the dataset. The nature of the factor models methodology are a perfect fit for the task, as they are essentially built for the purposes of data scrutiny and major trend identification. Statistical factor models can be applied to all data sets that have a factor structure, and moreover, their generality and ability to rationalise seemingly random data has not gone unnoticed in empirical research. Over the past fifteen years factor models methodology has been increasingly related to macroeconomic analysis and research into forecasting, interest rates and inflation studies, monetary policy, nowcasting procedures and a plethora of trading methodologies. Rudebusch and Wu (2008) developed a macrofinance model based on large dimensional factor models of an array of macroeconomic factors, providing an indicator of a country's economic health. Eickmeier and Ziegler (2008) tested the strength of factor models in the forecasting output and inflation of the countries. Bernanke and Bovin(2000) demonstrated applicability of the factor model methodology to the monetary policy identification. All these papers demonstrate the generality of the factor models approach, their ability to extract major trends, which can in turn be further interpreted in economic and financial terms by the application of the factor rotation methodology and evaluation of association between factors and time-series variables in the dataset (see Connor and Korajczyk (2009)).

The applicability of factor models to a wide variety of economic and financial problems results in the great importance of the topics related to the estimations of factor models. Indeed, one would where pos-

sible apply a consistent estimation technique, resulting in an unbiased, efficient and consistent factor. In other words, estimators should converge to the true common factors of the dataset; given the convergence is achieved then we can treat estimated factors as known and proceed with underlying research. The purpose of this chapter is to describe an addition to the existing factor models theory which allows faster rates of convergence and closer approximation of the true common factors. The benefits of the research can be seen in both theoretical and empirical econometrics. From a theoretical prospective the research contributes to the existing theory of factor models and proves the theoretical concept: that better rates of convergence can be achieved. At the same time the assumptions used in the theoretical proofs are similar to the classical assumptions of Stock and Watson (2002), Bai(2003), Bai(2004). It implies that under similar assumptions and without additional strengthening of underlined assumptions, we are able to propose a methodology that deliver more consistent and robust factor trends.

From an empirical econometrics prospective the research provides alternative generalized methodology, permitting greater precision in the estimation of common trends, without superimposing the stronger assumptions of the existing theory. Therefore, the research can be applied to a variety of empirical areas, and is especially valuable when the additional degree of the precision is necessary. We can identify a few areas, such as spread trading research (especially on the markets with tighten spreads) where the degree of the accuracy n identifying market entry points; the same applies to high frequency factor models research for the purpose of building trading strategies. Macroeconomic research which applies factor model theory will benefit from the more precise identification of the factors, however the impact would be less noticeable as macroeconomic research usually focusses on the identification of only generalized trends and indicators of economic development.

In this chapter we demonstrate that our methodology iemphasizes the robustness of estimated common components, in comparison to the case when stationary factors are directly estimated using first-differenced data. Improvement in estimations obtained due to the faster rates of convergence of higher order terms lead to better, which leads to more accurate finite sample properties than direct estimation from first-difference data. Using our findings we are able to complement the existing theory of factor models and contribute to the literature on factor models. We believe that our findings is important from a theoretical perspective, as for the past three decades the literature on factor models focusses on the development of the more robust and consistent estimators of large dimensional panel trends. In this respect our research is one in a long line of those papers focused on the problem of increasing robustness and consistency of the estimators. Therefore, the purpose of this chapter is to complete the existing

inferential theory for the stationary and non-stationary factor models by proposing a novel estimation methodology for common factors, loadings, and common components, in the context of large time series (T) and cross sectional (N) dimensions.

In addition, we contribute to the empirical research by providing methodology that insures higher theoretical precision of the factors. In empirical research we have to impose strong assumptions about data generating process in order to claim that factor trend of the large dimensional panel is consistent and robust. Our methodology provides a combination of the classical assumptions, and the alternative approach should the data generating process not satisfy the assumptions of the classical theories as described in Stock and Watson (2002) and Bai(2003). Additionally, given the generality of the assumptions, the methodology can be applied to a variety of the datasets so as to increase the precision of the factor estimator. This is especially valuable in the research which investigates the possibility of trading activities using real time ultra-high frequency data, or tight spread trading, where estimated trends are aimed at the development of future trading strategy and formation of the trading positions. As stated previously, the methodology is applicable to any research satisfying Stock and Watson(2002), Bai(2003), Bai(2004) classical factor models assumptions. However, using the methodology described in the chapter we are able to achieve more robust and consistent estimators than those described in the previous chapters.

The remainder of this chapter is organised as follows: chapter 2.2 gives a brief overview of the literature associated with the topic; chapter 2.3 provides the detailed technical specification of the model and describes the methodology; chapter 2.4 presents the theoretical findings; chapter 2.5 outlines the results of the simulation exercise; and chapter 2.6 remarks upon our conclusions.

3.2 Literature review

The paper concentrates on the inferential theory of static large dimensional factor models, to which we tailor the literature review. The extended version of the literature review in the area of factor models is outlined in the chapter 2.1. In this chapter we present an overview of the history of inferential theory development focused on static models.

Primary factor models have been extended to the large dimensional models by reducing the number of assumptions applied to the common factors and error terms. In the classical works of Lawley and Maxwell (1971), Anderson and Rubin (1956), and Anderson (1984), we find the original list of assumptions of primary factor models that had been modified. Among others, we referred to the set of assumptions

concentrating on cross-sectional and time-series independence of idiosyncratic errors; normality of the distribution of idiosyncratic errors as well as the factors; and the assumption that limits the number of column vectors in the panels.

The first transformation of primarily factor models took place using relaxation of the assumption regarding the independence of idiosyncratic terms, by allowing a weak cross-sectional correlation of errors as long as the dataset correlation matrix produces bounded largest eigenvalues. The transformation was first described in the work by Chamberlain and Rothschild (1983), who introduced the notion of approximate factor model. The work on development of approximate factor models continued, and in the late 1990's Stock and Watson introduced static large dimensional factor models. During the same period Forni, Hallin, Lippi, and Reichlin proposed their specification of the large dimensional factor model that is commonly referred to as the dynamic factor model. Boivin and Ng (2005) performed a comparison between the models, concluding that when applied to econometrics, both in fact produced similar forecasting results.

Theoretical findings in the 1980's and 90's played a crucial role in the development of the theory of large dimensional factor models. Further research laid a foundation for the new field new field of approximate large dimensional factor models. "Large" refers to the infinite number of model observations, as well as column vectors, and 'approximate' indicates relaxed assumptions regarding independence relaxed assumptions regarding independence of the idiosyncratic errors.

In the factor models, the dataset X_{it} is the only observable part of the model; factor loading, common factors and idiosyncratic factors are latent, and they are not observed and do not have direct interpretation. The dataset X_{it} is infinite and thus, maximum likelihood estimation technique is not applicable, though ML is commonly applied to the classical factor models. To estimate latent components of large dimensional models it was necessary to develop a technique that ensured consistency and unbiased estimators.

Connor and Korajczyk (1986) were the first to suggest an estimation solution for the factor models with an infinite number of factors. Their methodology was originally developed for the Chamberlain and Rothschild (1983) the approximate factor model. We recall that approximate factor model allows only one dimension of the dataset to reach infinity. Connor and Korajczyk (1986) took advantage of this feature and suggested a Principal Component methodology, analysing a covariance matrix of the finite dimension of the dataset. Such a transformation ensured consistent estimation of the factors in the

approximate factor model with a weak cross-sectional dependence on error terms.

Stock and Watson (2002) extended the Principal Component methodology for the samples with both infinite time-series and cross-sectional dimensions. The methodology employs the quasi-maximum likelihood technique for principal component analyses, allowing the presence of cross-sectional and time-series correlations of error terms. Additionally, Stock and Watson's (2002) methodology helped to impose time-varying factor loadings. Bai(2003) offerred a theoretical justification for the technique, analysing the distribution of the factors and loadings, deriving asymptotic rates of convergence when T/N of the sample goes to infinity. Anderson and Vahid (2007) improved the principal component estimation technique by allowing jumps in the dataset, and used IV approach to correct a measurement error from jumps.

To use principal component analysis for approximation of large dimensional factor models we have to impose a set of assumptions on the latent factors. The common factor assumptions ensure no degeneration, and each factor has a unique contribution to the variance of panel dataset. Additionally, factors are allowed to be correlated across time. The set of idiosyncratic assumptions ensures that errors can have weak cross-sectional dependence, which is similar to Chamberlain and Rothschild's (1983) model, in which weak time-series correlation of errors is permitted, as well as heteroskedasticity.

Further developments are found in Heaton and Solo (2006) who proved the robustness of the common factor estimators when the rate of cross-correlation increases with the speed of N. This development significantly improves the Chamberlain and Rothschild's (1983) original assumption regarding cross-sectional errors. Bai(2003) specified that another assumption of the original model can be relaxed by allowing weak correlation between common factors and idiosyncratic terms. Chapter 2.3.1 gives a rigorous examination of the factor models assumption. As long as the set of idiosyncratic assumptions holds, the estimators of common factor and loading is consistent. The static factor model also includes a set of strong assumptions, including the independence of common factors and factor loading, as the independence of loading and error terms.

Bai and Ng(2002) developed a formal methodology to estimate the optimal number of parameters in the large dimensional approximate factor model. They developed an information criteria methodology that is based on the filtration of the largest eigenvalues in the sample. Additionally, the number of optimal parameters can be tested using a scree plot. Onatski (2010) justified this methodology by providing asymptotic distribution for the factors extracted using scree plot tests. The author chooses optimal

number of factors by analysing the change in the slope and curvature of the scree plot of eigenvalues.

Stationary static large dimensional factor models experienced a growth in popularity amongst empirical studies. Numerous practitioners apply this approach to achieve consistent evaluation of large datasets. Among others are IV-factor model Kapetanios and Marcellino (2006a), Linear Factor Augmented Regressions by Bai and Ng (2006a) and a groundbreaking work of Stock and Watson (2002a) on Diffusion Index (DI) forecasting methodology.

In addition to the static factor models, there has been a growing discussion regarding the application of non-stationary models, allowimg for the analysis of the non-stationary set of variables Xit. The common factor of the non-stationary set of variables has a unit root. Bai (2004) developed an asymptotic theory and the rates of convergence for the non-stationary common factor. The theory holds when both time series observations and cross-sectional columns are infinite, leading to uniformly distributed common components. Moreover, the common factor is consistently estimated even if each error term is spurious. This is a significant advantage in comparison to the traditional multiple regressions where number of variables is limited. Bai (2004) imposes a number of assumptions.

In this chapter we give a brief desciption of the main results, before giving a full set of assumptions in chapter 2.3.1. The first set of assumptions imposes the condition that variance matrix of common factor, which has to be positive definite. Next, the non-stationary model implies similar conditions on the idiosyncratic term, such that they are allowed to be weakly cross sectional and time series dependent. The heteroskedasticity is allowed, whereas the dependence between errors and common factors is not. To detect the non-stationarity of the variables in the model, Bai and Ng (2002) developed PANIC (Panel Analysis of Non-stationarity in Idiosyncratic and Common Components), which examines the unobservable latent factors and demonstrates the number of stochastic trends that are driving the data. The majority of the unit root tests examine the dataset to determine a unit root of the trend, differentiating a PANIC approach.

The literature review provides a general overview of the main contributions of the theory of factor models. This paper concentrates on the theoretical findings, and we threfore amend the discussion of empirical applications of the large dimensional static factor models. The overview of the empirical developments in the field of large dimensional factor models can be found in the following works of Stock and Watson (2002b), Artis et al. (2005), Marcellino et al. (2003), Schumacher (2012). Unique developments in the empirical applications of factor models can be found in Bernanke and Boivin (2003),

Giannone et al. (2005a, b), Favero et al. (2005), Stock and Watson (2005).

3.3 Methodology

Consider the non-stationary panel factor series

$$X_{it} = \lambda_i' F_t + e_{it}, \tag{5}$$

where $i = 1, ..., n$, $t = 1, ..., T$, F_t is a k-dimensional vector whose DGP is assumed to be $F_t = F_{t-1} + \varepsilon_t$; in addition, we assume that e_{it} is stationary. Bai (2004) develops the inferential theory for (5) - specifically, for F_t, λ_i, and for the non-stationary common component $C_{it} \equiv \lambda_i' F_t$. On the other hand, one may also consider the stationary, first-differenced panel factor model

$$x_{it} = \lambda_i' f_t + u_{it}, \tag{6}$$

where $x_{it} = \Delta X_{it}$ and $f_t = \Delta F_t$. In this case, estimators for λ_i, f_t and $c_{it} \equiv \lambda_i' f_t$ (say $\hat{\lambda}_i$, $\hat{f}_t$ and $\hat{c}_{it}$ respectively) are provided by Bai (2003).

The purpose of this note is to complement the existing inferential theory on (5) and (6), by providing some results on estimation based on using the first difference of the estimator of F_t, say $\hat{F}_t$, computing (5). Indeed, instead of estimating f_t from (6), one could think of using $\tilde{f}_t = \hat{F}_t - \hat{F}_{t-1}$. Therefire, using either the estimation λ_i from (5), say $\hat{\lambda}_i$, or estimating λ_i from (6) using $\tilde{f}_t$, one can compute the first differenced estimator of c_{it} as $\tilde{c}_{it} \equiv \tilde{\lambda}_i' \tilde{f}_t$. Estimating f_t and c_{it} is useful for various purposes, and one important example is the estimation of the long run covariance matrices (henceforth, LRV) of F_t and C_{it}. Of course, this can be also done by using other techniques, such as the estimation of the LRV of C_{it},which can be achieved directly, using X_{it}; the LRV of F_t can be estimated using $\hat{f}_t$, calculated from (6). In this note, we consider the estimation based on $\tilde{f}_t$ and $\tilde{c}_{it}$.

In the context of bootstrapping nonstationary factor models, some results have already been developed by Trapani (2012a, 2012b). This note completes the inferential theory of the first-differenced estimators. In particular, in Section 3.4, we report rates of convergence for: $\tilde{f}_t$; for the estimator of λ_i based on using $\tilde{f}_t$ in (6), say $\tilde{\lambda}_i$; and for a weighted-sum-of-covariances estimator of the LRV of C_{it} based on $\tilde{f}_t$.

3.4 Theoretical Results

All results reported here for $\tilde{f}_t$, $\tilde{c}_{it}$ and $\tilde{\lambda}_i$ are derived under the same assumptions as in Bai (2003, 2004), which we omit for brevity. Henceforth, we define the $r \times r$ rotation matrix $H \equiv \left(\frac{\hat{F}'F}{T^2}\right)\left(\frac{\Lambda'\Lambda}{n}\right)$, where $F = [F_1, ..., F_T]'$ ($\hat{F}$ is defined similarly) and $\Lambda = [\lambda_1, ..., \lambda_n]'$. The number of factors, r, is assumed known from simplicity.

Firstly, we demonstrate a Lemma containing rates of convergence for $\tilde{f}_t = \hat{F}_t - \hat{F}_{t-1}$.

Lemma 1 *As $(n,T) \to \infty$, it holds that*

$$\tilde{f}_t - H'f_t = O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{1}{T^{3/2}}\right), \tag{7}$$

$$\max_{1\leq t\leq T}\left\|\tilde{f}_t - H'f_t\right\| = O_p\left(\frac{1}{T}\right) + O_p\left(\sqrt{\frac{T}{n}}\right), \tag{8}$$

$$\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{f}_t - H'f_t\right)u_{it} = O_p\left(\frac{1}{\sqrt{n}}\right) + O_p\left(\frac{1}{T^{3/2}}\right). \tag{9}$$

Under $\frac{n}{T^3} \to 0$, $\sqrt{n}\left(\tilde{f}_t - H'f_t\right) \xrightarrow{d} QN(0, \Upsilon_t)$, where Q is defined in Theorem 2 in Bai (2004, p. 148) and $\Upsilon_t \equiv \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} E\left(\lambda_i\lambda_j' u_{it}u_{jt}\right)$.

Lemma 1 states that rates and uniform convergence of $\tilde{f}_t - H'f_t$ are the same as for $\hat{F}_t - H'F_t$. This can be compared with the results in Theorem 1 in Bai (2003), where it is shown that $\hat{f}_t - H_1'f_t = O_p\left(n^{-1/2}\right) + O_p\left(T^{-1}\right)$ - note that, in general, the rotation matrices H and H_1 are different. Therefore, heuristically, $\tilde{f}_t$ should be a better estimator than $\hat{f}_t$ for the space spanned by f_t, especially when T is small. Lemma 1 is a complement, regarding the properties of $\tilde{f}_t$, to Lemma A.1 in Trapani (2012a; see Trapani 2012b for proofs). The Lemma contains essentially technical results that are useful for proofs.

We now turn to the estimation results of the loadings λ_i. To this end, it is possible to use the estimator of λ_i from (5), say $\hat{\lambda}_i$. Bai (2004, p. 148-149) shows that $\hat{\lambda}_i$ is "superconsistent", viz. $\hat{\lambda}_i - H^{-1}\lambda_i = O_p\left(T^{-1}\right)$; note also that the rate of convergence does not depend on n. Alternatively, it is possible to estimate loadings using $\tilde{f}_t$, by defining $\tilde{\lambda}_i = \left[\sum_{t=1}^{T}\tilde{f}_t\tilde{f}_t'\right]^{-1}\left[\sum_{t=1}^{T}\tilde{f}_t x_{it}\right]$. Let $\Sigma_\varepsilon \equiv E\left(\varepsilon_t\varepsilon_t'\right) = E\left(f_t f_t'\right)$; it holds that:

Proposition 1 *As $(n,T) \to \infty$ it holds that $\tilde{\lambda}_i - H^{-1}\lambda_i = O_p\left(n^{-1}\right) + O_p\left(T^{-1/2}\right)$. Under $\frac{\sqrt{T}}{n} \to 0$, $\sqrt{T}\left(\tilde{\lambda}_i - H^{-1}\lambda_i\right) \xrightarrow{d} N(0, V_i)$ with $V_i = \left(H'\Sigma_\varepsilon H\right)^{-1}\left(H'\Phi_i H\right)\left(H\Sigma_\varepsilon H'\right)^{-1}$ and $\Phi_i = \lim_{T\to\infty} E\left(f_t f_s' u_{it}u_{is}\right)$.*

Proposition 1 states that the properties of $\tilde{\lambda}_i$ are (modulo the rotation matrix H which is different from the case of using stationary data) the same as discussed in Bai (2003), where the estimation of λ_i is based on using (6). The result can be compared with $\hat{\lambda}_i$, whose convergence rate does not depend on n and it is faster in T.

Based on Lemma 1 and Proposition 1, consider the first-differenced estimator of the common components c_{it}, $\tilde{c}_{it} \equiv \hat{\lambda}_i' \tilde{f}_t = \hat{C}_{it} - \hat{C}_{it-1} = \hat{\lambda}_i' \left(\hat{F}_t - \hat{F}_{t-1} \right)$. By combining the results above, and using Lemma 3 in Bai (2004), we have $\tilde{c}_{it} - c_{it} = \hat{\lambda}_i' \tilde{f}_t - \lambda_i' f_t = \left(\hat{\lambda}_i - H^{-1} \lambda_i \right)' \tilde{f}_t + \left(\tilde{f}_t - H' f_t \right)' H^{-1} \lambda_i + \left(\hat{\lambda}_i - H^{-1} \lambda_i \right)' \left(\tilde{f}_t - H' f_t \right) = O_p \left(n^{-1/2} \right) + O_p \left(T^{-1} \right)$. In view of this, and using Theorem 3 in Bai (2004) on the limiting distribution of $T \left(\hat{\lambda}_i - H^{-1} \lambda_i \right)$, the asymptotic distribution of $\tilde{c}_{it} - c_{it}$ has the same properties as in Theorem 4 in Bai (2004, p. 149).

The results in Lemma 1 and Proposition 1 can be combined in order to estimate the LRV of the common factors F_t and of the common components C_{it}. Let Σ_F be the LRV of F_t, and define similarly the LRV of C_{it} as Σ_C. A possible way of estimating (a rotation of) Σ_F is attained through

$$\hat{\Sigma}_F = \hat{\gamma}_0^F + \sum_{j=1}^{h} \left(1 - \frac{j}{h+1} \right) \left(\hat{\gamma}_j^F + \hat{\gamma}_j^{F\prime} \right),$$

where h is a bandwidth parameter and $\hat{\gamma}_j^F \equiv T^{-1} \sum_{t=j+1}^{T} \tilde{f}_t \tilde{f}_{t-j}'$; other kernels can also be employed. This estimator is expected to be consistent in some sense under standard assumptions, on the decay of the autocorrelation coefficients of f_t. Of course, $\hat{\Sigma}_F$ does not estimate Σ_F consistently due to rotational indeterminacy; it can be expected that $\left\| \hat{\Sigma}_F - H' \Sigma_F H \right\| = o_p(1)$.

Similarly, Σ_C can be estimated either as $\hat{\Sigma}_C = \hat{\lambda}_i' \hat{\Sigma}_F \hat{\lambda}_i$, or as $\tilde{\Sigma}_C = \tilde{\lambda}_i' \hat{\Sigma}_F \tilde{\lambda}_i$. By virtue of Proposition 1, $\hat{\Sigma}_C$ should be better, and we focus our attention on it.

Theorem 1 *Assume that* $\sum_{j=0}^{\infty} j^s \left| \gamma_j^F \right| < \infty$. *It holds that*

$$\left\| \hat{\Sigma}_C - \Sigma_C \right\| = O_p \left(\frac{h}{\sqrt{T}} \right) + O_p \left(\frac{h}{n} \right) + O_p \left(\frac{1}{h} \right). \tag{10}$$

Theorem 1 contains rates of convergence for $\hat{\Sigma}_C$, which is a consistent estimator provided that $h \to \infty$ and that $h / \min \left\{ n, \sqrt{T} \right\} \to 0$. This also gives a selection rule for h; as an example, the choice of the bandwidth that maximizes the speed of convergence is $h^* = O \left(\min \left\{ T^{1/4}, n^{1/2} \right\} \right)$.

We point out that $\hat{\Sigma}_C$ is not the only possible estimator for Σ_C. As another factor-based alternative, one could consider estimating a rotation of Σ_F using $\hat{f}_t$ calculated from (6). Given that the rotation matrix H differs depending on whether (5) or (6) is used, in this case it is necessary to employ the estimated loadings from model (6), which has the same properties as $\tilde{\lambda}_i$ in Proposition 1. Based on this, and on Lemma 1, it can be expected that this estimator does not converge as fast as $\hat{\Sigma}_C$. Similarly, it is possible to use a weighted-sum-of-covariance estimator for Σ_C based on using the x_{it}s directly. Theoretically, this estimator should work due to the e_{it}s in (5) being stationary, although this may introduce some noise in the estimation of Σ_C.

3.5 Simulation Results

We report a Monte Carlo exercise to illustrate the behavior of the estimated common factors $\tilde{f}_t$ and common components $\tilde{c}_{it}$ in comparison to true counterparts. We perform simulation using large sample size matrixes. Although our paper addresses the problems related to the large sample size data, we felt compelled by the idea to experiment with the smaller sample of data; the results of the experiment are compared with the large sample experiment. The DGP we employ is the same as in Bai (2003), and in particular the error term in (5), e_{it}, is simulated according to ARMA(1,1) process, viz. $e_{it} = \rho e_{it-1} + v_{it} + \theta v_{it-1}$, with v_{it} i.i.d. standard normal. We report results for the following combinations of $(\rho, \theta) = \{(0,0), (0.75,0), (0,0.75), (0,-0.75), (0.5,0), (0,0.5), (0.-0.5)\}$; the combinations of the large sample size dimensions are as follow $(n,T) = \{25, 50, 100, 200, 500, 1000, 2000\} \times \{50, 100, 1000\}$; additionally we experiment with the finite sample of the following dimensions $(n,T) = \{5, 10\} \times \{5, 10, 15\}$. Number of replications is set to 1000.

We start the evaluation of the methodology with an assessment of the correlation parameters between common factors and true common factors simulated in the exercise; tables I, II, V, VI, IX and X report the results of the exercise. Tables I, V and IX relay the results for the exercise performed using the methodology described in Bai(2003), while tables II, VI and X demonstrate the results of the novel methodology.

The results suggest that the common factor estimated using the novel methodology bares greater correlation to the true factor than the traditional common factor. This results hold for all variations of the simulations, and therefore, the simulation results confirm our theoretical findings that the novel methodology provides faster rates of convergence and better approximation results. Coefficients have

better approximation when the number of column vectors increases in the panel dataset. We can observe a similar dynamic amongst by increasing time-series observations. The largest datasets have marginal deviation between the estimated and the true factor. As we decrease the numbers in column N and the time-series observations T, the correlation between true and estimated factors decreases. Bai(2003) described similar results which lead to the conclusion that larger panels of data have better approximations of estimated factors.

We simulate error components according to four separate processes. First we generate an error term as an iid process $ARMA(0,0)$, where errors are normally distributed. Second, DGP is generated using autocorrelation process $ARMA(p,0)$ where structural parameter p equals 0.5 and 0.75. The third and fourth processes are generated in a similar way to the moving average process $ARMA(0,q)$where q parameter can take values $\{0.5, 0.75, -0.5, -0.75\}$. We start from only one parameter p and q in the $ARMA(p,q)$ DGP for error terms. Next we increase the number of p and q parameters in the $ARMA$ process. Diversity in our simulations secures a robust evaluation of the new methodology results and helps to examine the impact on the long-run variance of the estimators.

All simulation exercises are able to extract common factors which are highly correlated to the true factors. The data set with errors following the iid process generated by using $ARMA(0,0)$ equation for the error terms demonstrated the best results. The majority of datasets with larger than $N = 25$ demonstrate correlation above 70% between true and estimated common components. If we increase the dataset to 100 column vectors than the correlation increases above 90%. When the number of variables is increased to 2000, the correlation between estimated factors and true parameters is exceptionally high, and on average stands at 99%. For extra large samples, with the number of variables N larger than 1000, our simulation demonstrates only marginal differences with classical methodology. However, panels N lower than 200 suggest that common factors extracted using novel methodology have higher correlation with true factors. Panels with a smaller sample size demonstrate wider dispersion of the results. However, the dynamic remains similar to the large sample panels. The correlation between estimated and true common factors increases with the number of column vectors. The results are mixed with the increased time- series observations, however when $T = 1000$ the results consistently improve in comparison to the smaller samples. Positive results of the simulation exercise suggest that the proposed methodology reduces the bias in long-run variance estimators and increases the convergence of the common factors.

A diverse number of autocovariates in the error term helps to address the issue of bandwidth using Monte-Carlo simulation. Theorem 1 suggests that the rates of convergence of estimated factors are better

than in Bai's (2003) methodology given that the bandwidth h goes to infinity. We test this assumption by changing the number of covariates of error terms and estimating the degree of convergence using the correlation coefficient as the indicator. Our results report that in practice the degree of variations induced by changing covariates is minimal and thus even a small bandwidth will demonstrate good results.

The dynamic of the common components is similar to common factors. We notice that the degree of correlation between common components is overall lower than among common factors. We explain this phenomenon by the fact that common components are the product of the multiplication of factor loadings and factor trends; factor loadings have a higher degree of variability which increases the dispersion of common factors. We can observe that the dynamics of common components and variability is similar to the dynamics of common factors. The correlation increases with expanding the number of variables N. For the smaller samples the correlation between true and estimated components is around 70% on average, although the standard deviation is relatively high. When we increase the number of N to 2000 the correlation rises to more than 80% on average. We notice similar patterns between correlations of common factors and common components, which we attribute to the fact that common components include common factors.

The correlation between smaller sample common components demonstrates similar dynamics. Overall, the correlation coefficients are significantly lower for smaller samples with an average of 50%. The component estimated the using new methodology demonstrated a higher correlation than the traditional methodology. Based on the results of the empirical exercise we can conclude that the new methodology has a positive effect on the precession of the estimation of common factors and components. Larger sample panels help estimators to converge to true value with faster rates. The diversity of the samples and equations provides a guide to a LRV, demonstrating increased correlation with an increased sample size. Our evaluation of a number of autocovariates in equations demonstrates that the results hold, given a different bandwidth level. Overall, the novel methodology demonstrates a smoothening of the results with increased estimator precision.

3.6 Conclusion

In this chapter we present an augmentation of the existing asymptotic theory of factor models. The aim of our research is to present a novel methodology which aims to enhance the rates of convergence for the estimated factor trends and common components, without additional assumptions. Our paper is motivated by the trend in the factor model literature, which has over the past 30 years moved towards closer approximation of the true common factors by estimators given more relaxed assumptions. The benefits of the research can be seen in both theoretical and empirical econometrics. From a theoretical prospective the research contributes to the existing theory of factor models and proves that better rates of convergence can be achieved. Simultaneously, the assumptions used in the theorised proofs are similar to the classical assumptions of Stock and Watson (2002), Bai(2003) and Bai(2004). It implies that under similar assumptions and without additional strengthening of underlined assumptions, we are able to propose a methodology which delivers more consistent and robust factor trends.

Our research is the result of careful investigation of the existing literature on factor models, where factor models inference distinguishes between stationary and non-stationary datasets Xit. In the proposed methodology we combine the results of the existing literature and develop a method that allow better approximation of the latent parts of factor the model. The methodology is based on the idea of estimation of the latent components of the factor models from the non-stationary panels. The appendix contains extensive proofs, on the basis of which we suggest that asymptotic characteristics of the latent factors is better when we estimate first difference factors from common factors in levels. The classical theory estimates first-difference factors by applying principle components analyses to the first-difference dataset. The methodology has the same first order terms in comparison to the existing estimators, but offers different asymptotic results for the higher order terms. As a result, estimators computed according to our novel methodology have a higher rate of convergence and provide a more robust approximation of the model parameters.

We convey a Monte-Carlo simulation to access a performance of the theoretical results. All simulations demonstrate,in comparison to the classical approach, better convergence of the estimated common factor when use novel methodology. We see that extremely large factor models (where the number of parameters is close to 2000) display only marginal deviation from the true factor. The results of the current methodology and classical approach are almost identical for extra large panels. The main difference is seen in the inference of smaller panels, where the number of column vectors is between 5 and 50. The same dynamics are observed in the common component simulations. We can see that common

components estimated using the new methodology are marginally closer to the true components; further more, the correlation increases with sample size.

In the economic and financial literature this type of model is most commonly used for the research proposes. We can see that novel methodology consistently produce higher correlation coefficients between true and estimated factors. These results confirm superiority of the novel methodology, in particular for smaller empirical datasets. Larger datasets demonstrate only marginal improvements during applications of the methodology. The methodology does not have any limitations, and can be applied to any factor model dataset, which makes it relevant not only from a theoretical prospective, but also for empirical application.

AppendixA

Table I Correlation Coefficient between $\hat{f}_t$ and f_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p=0, q=0$	T=50	0.8652	0.9256	0.9601	0.9845	0.9887	0.9926	0.9931
	T=100	0.8791	0.9368	0.9676	0.9754	0.9888	0.9856	0.9937
	T=1000	0.9690	0.9700	0.9755	0.9771	0.9895	0.9945	0.9985
$p=0.75, q=0$	T=50	0.9277	0.9654	0.9826	0.9890	0.9893	0.9976	0.9923
	T=100	0.9333	0.9645	0.9806	0.9703	0.9973	0.9981	0.9994
	T=1000	0.9609	0.9664	0.9890	0.9811	0.9910	0.9921	0.9994
$p=0, q=0.75$	T=50	0.5970	0.7520	0.7847	0.9906	0.9882	0.9985	0.9929
	T=100	0.7421	0.8578	0.9224	0.9774	0.9867	0.9823	0.9859
	T=1000	0.8455	0.8627	0.9209	0.9858	0.9958	0.9889	0.9922
$p=0, q=-0.75$	T=50	0.7350	0.8430	0.9192	0.9869	0.9933	0.9945	0.9996
	T=100	0.7947	0.9426	0.9450	0.9680	0.9975	0.9900	0.9901
	T=1000	0.8953	0.9512	0.9635	0.9915	0.9961	0.9976	0.9986
$p=0.5, q=0$	T=50	0.9079	0.9141	0.9086	0.9839	0.9870	0.9955	0.9992
	T=100	0.9056	0.9591	0.9731	0.9753	0.9801	0.9889	0.9899
	T=1000	0.9681	0.9629	0.9695	0.9870	0.9891	0.9958	0.9989
$p=0, q=0.5$	T=50	0.8665	0.8910	0.9481	0.9827	0.9828	0.9995	0.9939
	T=100	0.8677	0.9010	0.9494	0.9732	0.9784	0.9896	0.9937
	T=1000	0.9071	0.9602	0.9594	0.9732	0.9790	0.9948	0.9980
$p=0, q=-0.5$	T=50	0.8307	0.9215	0.9496	0.9774	0.9853	0.9862	0.9911
	T=100	0.8401	0.9399	0.9649	0.9889	0.9806	0.9917	0.9950
	T=1000	0.9328	0.9416	0.9628	0.9865	0.9893	0.9904	0.9931

Table I shows estimators of correlation coefficients between factors $\hat{f}_t$ estimated using a new methodology and the true factors f_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table II Correlation Coefficient between $\tilde{f}_t$ and f_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p=0, q=0$	T=50	0.8966	0.9435	0.9709	0.9855	0.9882	0.9969	0.9965
	T=100	0.8964	0.9442	0.9718	0.9753	0.9895	0.9973	0.9971
	T=1000	0.9787	0.9843	0.9871	0.9899	0.9945	0.9921	0.9983
$p=0.75, q=0$	T=50	0.9278	0.9655	0.9786	0.9920	0.9918	0.9977	0.9986
	T=100	0.9335	0.9652	0.9811	0.9858	0.9978	0.9982	0.9875
	T=1000	0.9750	0.9789	0.9824	0.9904	0.9917	0.9930	0.9995
$p=0, q=0.75$	T=50	0.8381	0.9098	0.9586	0.9973	0.9909	0.9954	0.9969
	T=100	0.8461	0.9148	0.9550	0.9591	0.9952	0.9976	0.9912
	T=1000	0.9473	0.9446	0.9490	0.9875	0.9842	0.9921	0.9932
$p=0, q=-0.75$	T=50	0.8192	0.9213	0.9577	0.9984	0.9836	0.9950	0.9988
	T=100	0.9084	0.9567	0.9548	0.9777	0.9912	0.9944	0.9995
	T=1000	0.9492	0.9606	0.9559	0.9603	0.9782	0.9930	0.9927
$p=0.5, q=0$	T=50	0.8417	0.9094	0.9623	0.9835	0.9800	0.9955	0.9994
	T=100	0.9175	0.9593	0.9740	0.9853	0.9882	0.9989	0.9937
	T=1000	0.9585	0.9659	0.9667	0.9890	0.9895	0.9975	0.9945
$p=0, q=0.5$	T=50	0.8783	0.9316	0.9689	0.9835	0.9883	0.9995	0.9992
	T=100	0.8766	0.9338	0.9455	0.9618	0.9882	0.9996	0.9907
	T=1000	0.9518	0.9612	0.9620	0.9862	0.9822	0.9916	0.9981
$p=0, q=-0.5$	T=50	0.9048	0.9431	0.9659	0.9780	0.9938	0.9962	0.9994
	T=100	0.9161	0.9441	0.9693	0.9890	0.9832	0.9916	0.9975
	T=1000	0.9436	0.9548	0.9635	0.9845	0.9812	0.9958	0.9965

Table II shows estimators of correlation coefficients between factors $\tilde{f}_t$ estimated using the classical methodology and the true factors f_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table III Correlation Coefficients between $\hat{c}_t$ and c_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p=0, q=0$	T=50	0.7461	0.7851	0.8241	0.8804	0.8476	0.8748	0.9018
	T=100	0.7297	0.8008	0.8239	0.8371	0.8823	0.8834	0.9090
	T=1000	0.8688	0.8325	0.8325	0.8359	0.8557	0.8911	0.9132
$p=0.75, q=0$	T=50	0.7808	0.8470	0.8593	0.8571	0.8597	0.8635	0.9072
	T=100	0.7918	0.8342	0.8533	0.8314	0.8836	0.8615	0.9183
	T=1000	0.8477	0.8254	0.8674	0.8741	0.8566	0.8576	0.9398
$p=0, q=0.75$	T=50	0.5530	0.6074	0.6531	0.8627	0.8847	0.8676	0.9166
	T=100	0.5928	0.7334	0.7933	0.8502	0.8491	0.8521	0.9091
	T=1000	0.7175	0.7521	0.7906	0.8764	0.8864	0.8817	0.9229
$p=0, q=-0.75$	T=50	0.6199	0.7109	0.7978	0.8636	0.8568	0.8450	0.9035
	T=100	0.6897	0.8282	0.8309	0.8271	0.8887	0.8438	0.9168
	T=1000	0.7790	0.8150	0.8402	0.8535	0.8716	0.8891	0.9210
$p=0.5, q=0$	T=50	0.7763	0.7890	0.7866	0.8680	0.8773	0.8743	0.9078
	T=100	0.7812	0.8415	0.8405	0.8428	0.8578	0.8528	0.9208
	T=1000	0.8284	0.8272	0.8314	0.8640	0.8838	0.8472	0.9310
$p=0, q=0.5$	T=50	0.7188	0.7716	0.8456	0.8689	0.8418	0.8598	0.9055
	T=100	0.7335	0.7521	0.8271	0.8292	0.8518	0.8400	0.9207
	T=1000	0.7675	0.8197	0.8457	0.8415	0.8543	0.8851	0.9430
$p=0, q=-0.5$	T=50	0.6969	0.8116	0.8230	0.8548	0.8713	0.8611	0.9171
	T=100	0.7285	0.8211	0.8372	0.8764	0.8643	0.8714	0.9161
	T=1000	0.8186	0.8196	0.8375	0.8402	0.8647	0.8458	0.9283

Table III shows estimator of correlation coefficient between common component $\hat{c}_t$ estimated using a new methodology and the true common component c_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table IV Correlation Coefficients between $\check{c}_t$ and c_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p=0, q=0$	T=50	0.7591	0.8206	0.8507	0.8462	0.8851	0.8588	0.9067
	T=100	0.7467	0.8219	0.8274	0.8368	0.8547	0.8557	0.9042
	T=1000	0.8287	0.8801	0.8423	0.8667	0.8826	0.8520	0.9299
$p=0.75, q=0$	T=50	0.8074	0.8300	0.8410	0.8587	0.8715	0.8494	0.9180
	T=100	0.8276	0.8386	0.8593	0.8725	0.8530	0.8724	0.9164
	T=1000	0.8702	0.8338	0.8566	0.8875	0.8562	0.8658	0.9473
$p=0, q=0.75$	T=50	0.7374	0.7850	0.8401	0.8871	0.8650	0.8799	0.9300
	T=100	0.7128	0.8016	0.8518	0.8570	0.8952	0.8908	0.9182
	T=1000	0.8101	0.8084	0.8478	0.8551	0.8472	0.8785	0.9135
$p=0, q=-0.75$	T=50	0.7058	0.7802	0.8148	0.8502	0.8751	0.8571	0.9008
	T=100	0.7904	0.8375	0.8427	0.8415	0.8625	0.8565	0.9196
	T=1000	0.8143	0.8156	0.8121	0.8457	0.8605	0.8656	0.9258
$p=0.5, q=0$	T=50	0.7346	0.7883	0.8423	0.8532	0.8553	0.8644	0.9138
	T=100	0.8159	0.8246	0.8371	0.8551	0.8512	0.8699	0.9214
	T=1000	0.8378	0.8475	0.8594	0.8538	0.8572	0.8619	0.9381
$p=0, q=0.5$	T=50	0.7492	0.7868	0.8292	0.8609	0.8437	0.8510	0.9132
	T=100	0.7625	0.8113	0.8138	0.8580	0.8646	0.8695	0.9147
	T=1000	0.8223	0.8285	0.8357	0.8596	0.8544	0.8838	0.9427
$p=0, q=-0.5$	T=50	0.7963	0.8257	0.8395	0.8495	0.8830	0.8798	0.9121
	T=100	0.8112	0.8321	0.8522	0.8578	0.8628	0.8843	0.9187
	T=1000	0.8146	0.8170	0.8321	0.8641	0.8761	0.8546	0.9429

Table IV shows estimators of correlation coefficients between common component $\check{c}_t$ estimated using the classical methodology and the true common component c_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table V Correlation Coefficient between $\hat{f}_t$ and f_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p_1 = 0, p_2 = 0, q_1 = 0, q_2 = 0$	T=50	0.8787	0.9107	0.9679	0.9887	0.9863	0.9978	0.9904
	T=100	0.8790	0.9483	0.9634	0.9724	0.9852	0.9887	0.9909
	T=1000	0.9659	0.9693	0.9702	0.9701	0.9842	0.9962	0.9936
$p_1 = 0.75, p_2 = 0.75, q_1 = 0, q_2 = 0$	T=50	0.9208	0.9785	0.9794	0.9824	0.9861	0.9925	0.9931
	T=100	0.9392	0.9602	0.9826	0.9784	0.9942	0.9977	0.9915
	T=1000	0.9619	0.9666	0.9848	0.9884	0.9979	0.9906	0.9983
$p_1 = 0, p_2 = 0, q_1 = 0.75, q_2 = 0.75$	T=50	0.5981	0.7795	0.7864	0.9829	0.9805	0.9955	0.9950
	T=100	0.7516	0.8503	0.9238	0.9632	0.9804	0.9896	0.9843
	T=1000	0.8329	0.8639	0.9253	0.9842	0.9915	0.9937	0.9951
$p_1 = 0, p_2 = 0, q_1 = -0.75, q_2 = -0.75$	T=50	0.7364	0.8457	0.9101	0.9808	0.9982	0.9974	0.9977
	T=100	0.7989	0.9553	0.9404	0.9644	0.9903	0.9992	0.9971
	T=1000	0.8990	0.9513	0.9630	0.9972	0.9979	0.9949	0.9971
$p_1 = 0.5, p_2 = 0.5, q_1 = 0, q_2 = 0$	T=50	0.8918	0.9214	0.9730	0.9804	0.9835	0.9954	0.9984
	T=100	0.9081	0.9509	0.9747	0.9768	0.9848	0.9858	0.9998
	T=1000	0.9626	0.9612	0.9653	0.9673	0.9823	0.9861	0.9976
$p_1 = 0, p_2 = 0, q_1 = 0.5, q_2 = 0.5$	T=50	0.8675	0.8999	0.9459	0.9818	0.9828	0.9972	0.9979
	T=100	0.8771	0.9094	0.9419	0.9790	0.9763	0.9883	0.9990
	T=1000	0.9072	0.9538	0.9533	0.9765	0.9793	0.9975	0.9920
$p_1 = 0, p_2 = 0, q_1 = 0, q_2 = 0$	T=50	0.7042	0.6839	0.7718	0.9476	0.9716	0.8647	0.8967
	T=100	0.7036	0.7323	0.8858	0.9930	0.95622	0.9607	0.9500
	T=1000	0.7189	0.7241	0.9561	0.9947	0.87572	0.9097	0.9586

Table V shows estimators of correlation coefficients between factors $\tilde{f}_t$ estimated using a classical methodology and the true factors f_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table VI Correlation Coefficient between $\tilde{f}_t$ and f_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p_1 = 0, p_2 = 0, q_1 = 0, q_2 = 0$	T=50	0.8996	0.9568	0.9727	0.9737	0.9888	0.9963	0.9965
	T=100	0.8953	0.9305	0.9769	0.9797	0.9899	0.9927	0.9958
	T=1000	0.9761	0.9767	0.9811	0.9879	0.9901	0.9960	0.9990
$p_1 = 0.75, p_2 = 0.75, q_1 = 0, q_2 = 0$	T=50	0.9248	0.9670	0.9742	0.9980	0.9998	0.9960	0.9987
	T=100	0.9378	0.9760	0.9805	0.9874	0.9917	0.9904	0.9903
	T=1000	0.9734	0.9756	0.9855	0.9957	0.9922	0.9916	0.9915
$p_1 = 0, p_2 = 0, q_1 = 0.75, q_2 = 0.75$	T=50	0.8325	0.9101	0.9542	0.9847	0.9904	0.9977	0.9964
	T=100	0.8453	0.9159	0.9523	0.9554	0.9979	0.9954	0.9989
	T=1000	0.9434	0.9431	0.9586	0.9737	0.9895	0.9938	0.9941
$p_1 = 0, p_2 = 0, q_1 = -0.75, q_2 = -0.75$	T=50	0.8180	0.9225	0.9516	0.9916	0.9830	0.9982	0.9927
	T=100	0.9148	0.9526	0.9678	0.9706	0.9950	0.9935	0.9954
	T=1000	0.9456	0.9675	0.9511	0.9664	0.9714	0.9902	0.9991
$p_1 = 0.5, p_2 = 0.5, q_1 = 0, q_2 = 0$	T=50	0.8479	0.9137	0.9673	0.9817	0.9837	0.9933	0.9969
	T=100	0.9109	0.9507	0.9667	0.9846	0.9889	0.9971	0.9908
	T=1000	0.9586	0.9651	0.9607	0.9662	0.9806	0.9983	0.9937
$p_1 = 0, p_2 = 0, q_1 = 0.5, q_2 = 0.5$	T=50	0.8813	0.9363	0.9785	0.9859	0.9897	0.9978	0.9926
	T=100	0.8723	0.9362	0.9588	0.9775	0.9888	0.9954	0.9966
	T=1000	0.9566	0.9668	0.9675	0.9802	0.9801	0.9934	0.9952
$p_1 = 0, p_2 = 0, q_1 = -0.5, q_2 = -0.5$	T=50	0.9064	0.9419	0.9635	0.9720	0.9911	0.9924	0.9951
	T=100	0.9231	0.9402	0.9662	0.9782	0.9860	0.9902	0.9987
	T=1000	0.9476	0.9522	0.9767	0.9846	0.9898	0.9917	0.9942

Table VI shows estimators of correlation coefficient between factors $\tilde{f}_t$ estimated using the classical methodology and the true factors f_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table VII Correlation Coefficients between $\hat{c}_t$ and c_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p_1 = 0, p_2 = 0, q_1 = 0, q_2 = 0$	T=50	0.7391	0.8171	0.8529	0.8698	0.8729	0.8538	0.9058
	T=100	0.7358	0.8177	0.8515	0.8568	0.8610	0.8812	0.9171
	T=1000	0.8517	0.8601	0.8611	0.8702	0.8893	0.8753	0.9085
$p_1 = 0.75, p_2 = 0.75, q_1 = 0, q_2 = 0$	T=50	0.8118	0.8484	0.8630	0.8589	0.8575	0.8686	0.9013
	T=100	0.8160	0.8211	0.8346	0.8593	0.8635	0.8772	0.9134
	T=1000	0.8312	0.8306	0.8455	0.8544	0.8827	0.8855	0.9398
$p_1 = 0, p_2 = 0, q_1 = 0.75, q_2 = 0.75$	T=50	0.5928	0.6336	0.7167	0.8431	0.8504	0.8674	0.9175
	T=100	0.6159	0.7154	0.8084	0.8520	0.8673	0.8882	0.9183
	T=1000	0.6981	0.7366	0.7910	0.8430	0.8839	0.8875	0.9150
$p_1 = 0, p_2 = 0, q_1 = -0.75, q_2 = -0.75$	T=50	0.5856	0.7123	0.7828	0.8450	0.8698	0.8636	0.9069
	T=100	0.6798	0.8260	0.8288	0.8345	0.8574	0.8631	0.9202
	T=1000	0.7628	0.8181	0.8553	0.8580	0.8760	0.8965	0.9270
$p_1 = 0.5, p_2 = 0.5, q_1 = 0, q_2 = 0$	T=50	0.7608	0.8092	0.8091	0.8637	0.8681	0.8835	0.9051
	T=100	0.7848	0.8507	0.8641	0.8636	0.8629	0.8806	0.9114
	T=1000	0.8191	0.8448	0.8466	0.8447	0.8554	0.8810	0.9290
$p_1 = 0, p_2 = 0, q_1 = 0.5, q_2 = 0.5$	T=50	0.7453	0.7546	0.8089	0.8426	0.8382	0.8953	0.9120
	T=100	0.7638	0.7720	0.8412	0.8485	0.8614	0.8632	0.9252
	T=1000	0.7644	0.8225	0.8386	0.8540	0.8334	0.8695	0.9464
$p_1 = 0, p_2 = 0, q_1 = -0.5, q_2 = -0.5$	T=50	0.7031	0.7725	0.8096	0.8753	0.8357	0.8547	0.9164
	T=100	0.7219	0.7997	0.8624	0.8793	0.8779	0.8737	0.9207
	T=1000	0.8031	0.8144	0.8355	0.8518	0.8678	0.8657	0.9219

Table VII shows estimators of correlation coefficient between common component $\hat{c}_t$ estimated using a new methodology and the true common component c_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using ARMA(p, q) type process.

Table VIII Correlation Coefficients between $\tilde{c}_t$ and c_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p_1 = 0, p_2 = 0, q_1 = 0, q_2 = 0$	T=50	0.7883	0.8355	0.8515	0.8554	0.8776	0.8727	0.9057
	T=100	0.7469	0.7852	0.8627	0.8759	0.8756	0.8873	0.9096
	T=1000	0.8594	0.8577	0.8613	0.8874	0.8741	0.8818	0.9328
$p_1 = 0.75, p_2 = 0.75, q_1 = 0, q_2 = 0$	T=50	0.8197	0.8333	0.8466	0.8647	0.8720	0.8749	0.9279
	T=100	0.8074	0.8704	0.8624	0.8742	0.8839	0.8860	0.9238
	T=1000	0.8360	0.8528	0.8486	0.8699	0.8711	0.8871	0.9546
$p_1 = 0, p_2 = 0, q_1 = 0.75, q_2 = 0.75$	T=50	0.7115	0.8074	0.8075	0.8642	0.8618	0.8800	0.9357
	T=100	0.7321	0.8119	0.8350	0.8302	0.8765	0.8727	0.9213
	T=1000	0.8001	0.8131	0.8115	0.8423	0.8606	0.8778	0.9100
$p_1 = 0, p_2 = 0, q_1 = -0.75, q_2 = -0.75$	T=50	0.6734	0.8005	0.8299	0.8534	0.8816	0.8833	0.8938
	T=100	0.7952	0.8239	0.8201	0.8631	0.8461	0.8577	0.9183
	T=1000	0.7964	0.8388	0.8419	0.8599	0.8632	0.8780	0.9306
$p_1 = 0.5, p_2 = 0.5, q_1 = 0, q_2 = 0$	T=50	0.7122	0.7975	0.8181	0.8501	0.8606	0.8739	0.9093
	T=100	0.8093	0.8193	0.8503	0.8529	0.8630	0.8622	0.9196
	T=1000	0.8273	0.8370	0.8349	0.8525	0.8666	0.8962	0.9454
$p_1 = 0, p_2 = 0, q_1 = 0.5, q_2 = 0.5$	T=50	0.7691	0.8347	0.8377	0.8525	0.8673	0.8890	0.9208
	T=100	0.7343	0.7955	0.8381	0.8643	0.8542	0.8791	0.9118
	T=1000	0.8267	0.8520	0.8571	0.8724	0.8536	0.8452	0.9376
$p_1 = 0, p_2 = 0, q_1 = -0.5, q_2 = -0.5$	T=50	0.7971	0.8073	0.8135	0.8277	0.8604	0.8924	0.9114
	T=100	0.7960	0.8090	0.8596	0.8525	0.8794	0.8866	0.9141
	T=1000	0.8242	0.8496	0.8619	0.8484	0.8717	0.8759	0.9344

Table shows an estimator of correlation coefficient between common component $\tilde{c}_t$ estimated using classical methodology and true common component c_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels were constructed using equation 6 from simulated factors, loading and error terms; the error terms were simulated using ARMA(p, q) type process.

Table IX Correlation Coefficient between $\hat{f}_t$ and f_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.8785	0.9218	0.9641	0.9787	0.9873	0.9952	0.9968
$q_1 = 0, q_2 = 0, q_3 = 0$	T=100	0.8717	0.9444	0.9681	0.9693	0.9888	0.9994	0.9980
	T=1000	0.9643	0.9728	0.9639	0.9792	0.9804	0.9901	0.9902
$p_1 = 0.75, p_2 = 0.75, p_3 = 0.75,$	T=50	0.9346	0.9669	0.9849	0.9839	0.9832	0.9929	0.9968
$q_1 = 0, q_2 = 0, q_3 = 0$	T=100	0.9394	0.9773	0.9729	0.9776	0.9847	0.9967	0.9917
	T=1000	0.9617	0.9697	0.9838	0.9822	0.9991	0.9952	0.9949
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.6140	0.7575	0.9879	0.9855	0.9975	0.9904	0.9926
$q_1 = 0.75, q_2 = 0.75, q_3 = 0.75$	T=100	0.7450	0.8605	0.9226	0.9742	0.9810	0.9974	0.9932
	T=1000	0.8420	0.8656	0.9344	0.9705	0.9940	0.9913	0.9963
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.7448	0.8493	0.9153	0.9892	0.9968	0.9988	0.9987
$q_1 = -0.75, q_2 = -0.75, q_3 = -0.75$	T=100	0.8035	0.9379	0.9413	0.9603	0.9664	0.9915	0.9926
	T=1000	0.8940	0.9637	0.9694	0.9873	0.9941	0.9983	0.9978
$p_1 = 0.5, p_2 = 0.5, p_3 = 0.5,$	T=50	0.9065	0.9192	0.9602	0.9785	0.9881	0.9967	0.9992
$q_1 = 0, q_2 = 0, q_3 = 0$	T=100	0.9175	0.9577	0.9607	0.9735	0.9888	0.9932	0.9979
	T=1000	0.9533	0.9666	0.9710	0.9802	0.9804	0.9951	0.9943
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.8735	0.8948	0.9400	0.9789	0.9819	0.9985	0.9953
$q_1 = 0.5, q_2 = 0.5, q_3 = 0.5$	T=100	0.8686	0.9154	0.9457	0.9775	0.9707	0.9945	0.9941
	T=1000	0.9183	0.9686	0.9587	0.9731	0.9757	0.9948	0.9949
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.5847	0.7807	0.8657	0.9273	0.9783	0.8805	0.9544
$q_1 = -0.5, q_2 = -0.5, q_3 = -0.5$	T=100	0.6055	0.8033	0.8544	0.8950	0.9709	0.9075	0.9852
	T=1000	0.5735	0.7634	0.8676	0.9686	0.9065	0.9061	0.9553

Table IX shows estimators of correlation coefficients between factors $\tilde{f}_t$ estimated using the classical methodology and the true factors f_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table X Correlation Coefficient between $\tilde{f}_t$ and f_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
	T=50	0.9066	0.9401	0.9665	0.9819	0.9888	0.9924	0.9978
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.8945	0.9566	0.9793	0.9725	0.9842	0.9935	0.9922
$q_1 = 0, q_2 = 0, q_3 = 0$	T=1000	0.9720	0.9813	0.9859	0.9881	0.9960	0.9976	0.9917
	T=50	0.9351	0.9668	0.9729	0.9800	0.9930	0.9922	0.9955
$p_1 = 0.75, p_2 = 0.75, p_3 = 0.75,$	T=100	0.9454	0.9639	0.9810	0.9889	0.9964	0.9965	0.9988
$q_1 = 0, q_2 = 0, q_3 = 0$	T=1000	0.9723	0.9612	0.9872	0.9879	0.9955	0.9937	0.9953
	T=50	0.8443	0.9640	0.9726	0.9949	0.9908	0.9986	0.9919
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.8418	0.9646	0.9810	0.9993	0.9906	0.9947	0.9984
$q_1 = 0.75, q_2 = 0.75, q_3 = 0.75$	T=1000	0.9465	0.9734	0.9849	0.9841	0.9853	0.9925	0.9981
	T=50	0.8209	0.9408	0.9508	0.9918	0.9883	0.9945	0.9914
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.9180	0.9557	0.9522	0.9798	0.9907	0.9999	0.9993
$q_1 = -0.75, q_2 = -0.75, q_3 = -0.75$	T=1000	0.9443	0.9559	0.9555	0.9654	0.9701	0.9979	0.9915
	T=50	0.8438	0.9646	0.9625	0.9707	0.9866	0.9916	0.9960
$p_1 = 0.5, p_2 = 0.5, p_3 = 0.5,$	T=100	0.9163	0.9332	0.9763	0.9754	0.9842	0.9924	0.9937
$q_1 = 0, q_2 = 0, q_3 = 0$	T=1000	0.9536	0.9675	0.9603	0.9824	0.9889	0.9952	0.9959
	T=50	0.8713	0.9671	0.9449	0.9862	0.9842	0.9931	0.9986
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.8848	0.9386	0.9669	0.9624	0.9764	0.9928	0.9938
$q_1 = 0.5, q_2 = 0.5, q_3 = 0.5$	T=1000	0.9593	0.9301	0.9663	0.9874	0.9806	0.9999	0.9953
	T=50	0.9197	0.9474	0.9604	0.9757	0.9964	0.9959	0.9999
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.9123	0.9529	0.9636	0.9715	0.9880	0.9917	0.9935
$q_1 = -0.5, q_2 = -0.5, q_3 = -0.5$	T=1000	0.9490	0.9519	0.9699	0.9870	0.9869	0.9911	0.9983

Table X shows estimators of correlation coefficients between factors $\tilde{f}_t$ estimated using the classical methodology and the true factors f_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table XI Correlation Coefficients between $\hat{c}_t$ and c_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.7730	0.8128	0.8413	0.8579	0.8674	0.8759	0.9111
$q_1 = 0, q_2 = 0, q_3 = 0$	T=100	0.7632	0.8389	0.8628	0.8647	0.8745	0.8848	0.9173
	T=1000	0.8186	0.8215	0.8377	0.8561	0.8647	0.8869	0.9113
$p_1 = 0.75, p_2 = 0.75, p_3 = 0.75,$	T=50	0.8338	0.8468	0.8667	0.8573	0.8622	0.8864	0.9075
$q_1 = 0, q_2 = 0, q_3 = 0$	T=100	0.8304	0.8670	0.8499	0.8475	0.8795	0.8956	0.9056
	T=1000	0.8480	0.8681	0.8644	0.8781	0.8802	0.8852	0.9445
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.5965	0.7565	0.8529	0.8637	0.8678	0.8722	0.9271
$q_1 = 0.75, q_2 = 0.75, q_3 = 0.75$	T=100	0.6010	0.7321	0.8062	0.8703	0.8720	0.8814	0.9234
	T=1000	0.6956	0.7506	0.8144	0.8704	0.8709	0.8802	0.9195
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.6334	0.7285	0.7855	0.8423	0.8873	0.8722	0.9152
$q_1 = -0.75, q_2 = -0.75, q_3 = -0.75$	T=100	0.6882	0.8016	0.8156	0.8474	0.8343	0.8654	0.9124
	T=1000	0.7835	0.8202	0.8299	0.8651	0.8456	0.8715	0.9340
$p_1 = 0.5, p_2 = 0.5, p_3 = 0.5,$	T=50	0.7869	0.8095	0.8103	0.8355	0.8692	0.8940	0.8974
$q_1 = 0, q_2 = 0, q_3 = 0$	T=100	0.7852	0.8441	0.8481	0.8670	0.8742	0.8889	0.9109
	T=1000	0.8217	0.8334	0.8410	0.8690	0.8710	0.8531	0.9285
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.7437	0.7521	0.8008	0.8364	0.8730	0.8852	0.9101
$q_1 = 0.5, q_2 = 0.5, q_3 = 0.5$	T=100	0.7497	0.7854	0.8154	0.8425	0.8371	0.8633	0.9329
	T=1000	0.7885	0.8661	0.8523	0.8393	0.8605	0.8693	0.9367
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=50	0.4681	0.6407	0.7358	0.8034	0.8321	0.7586	0.9133
$q_1 = -0.5, q_2 = -0.5, q_3 = -0.5$	T=100	0.4635	0.6842	0.7525	0.7928	0.8366	0.7974	0.9110
	T=1000	0.4353	0.6351	0.7550	0.8324	0.7590	0.7724	0.9171

Table XI shows estimators of correlation coefficients between common component $\hat{c}_t$ estimated using a new methodology and the true common component c_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table XII Correlation Coefficients between $\tilde{c}_t$ and c_t

		N=25	N=50	N=100	N=200	N=500	N=1000	N=2000
	T=50	0.8035	0.8116	0.8473	0.8657	0.8769	0.8784	0.9063
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.7610	0.8474	0.8482	0.8530	0.8697	0.8822	0.9172
$q_1 = 0, q_2 = 0, q_3 = 0$	T=1000	0.8296	0.8475	0.8803	0.8742	0.8902	0.8980	0.9358
	T=50	0.8113	0.8336	0.8574	0.8533	0.8559	0.8637	0.9205
$p_1 = 0.75, p_2 = 0.75, p_3 = 0.75,$	T=100	0.8388	0.8479	0.8419	0.8546	0.8514	0.8632	0.9156
$q_1 = 0, q_2 = 0, q_3 = 0$	T=1000	0.8450	0.8427	0.8781	0.8722	0.8850	0.8850	0.9521
	T=50	0.7143	0.8269	0.8417	0.8862	0.8825	0.8902	0.9264
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.6933	0.8571	0.8604	0.8779	0.8708	0.8845	0.9258
$q_1 = 0.75, q_2 = 0.75, q_3 = 0.75$	T=1000	0.8352	0.8358	0.8339	0.8406	0.8593	0.8909	0.9187
	T=50	0.6820	0.8342	0.8459	0.8766	0.8717	0.8884	0.9035
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.7999	0.8489	0.8065	0.8759	0.8765	0.8820	0.9140
$q_1 = -0.75, q_2 = -0.75, q_3 = -0.75$	T=1000	0.8389	0.8134	0.8473	0.8571	0.8537	0.8885	0.9329
	T=50	0.7148	0.8174	0.8278	0.8665	0.8726	0.8749	0.9115
$p_1 = 0.5, p_2 = 0.5, p_3 = 0.5,$	T=100	0.7671	0.8329	0.8691	0.8605	0.8777	0.8844	0.9208
$q_1 = 0, q_2 = 0, q_3 = 0$	T=1000	0.8311	0.8350	0.8615	0.8742	0.8747	0.8810	0.9476
	T=50	0.7466	0.8564	0.8232	0.8565	0.8673	0.8911	0.9161
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.7450	0.7999	0.8536	0.8579	0.8633	0.8797	0.9193
$q_1 = 0.5, q_2 = 0.5, q_3 = 0.5$	T=1000	0.8402	0.7995	0.8314	0.8649	0.8597	0.8823	0.9429
	T=50	0.7834	0.8380	0.8529	0.8689	0.8917	0.8851	0.9060
$p_1 = 0, p_2 = 0, p_3 = 0,$	T=100	0.8032	0.8263	0.8329	0.8632	0.8777	0.8900	0.9223
$q_1 = -0.5, q_2 = -0.5, q_3 = -0.5$	T=1000	0.8276	0.8329	0.8340	0.8739	0.8880	0.8938	0.9437

Table XII shows estimators of correlation coefficients between common component $\tilde{c}_t$ estimated using the classical methodology and the true common component c_t. Correlations are calculated for large dimensional panels with N number of columns and T number of observations. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(p, q) type process.

Table XIII Simulation results for smaller sample size correlation of common factors $\hat{f}_t$ and f

		$p=0,\ q=0$		$p=0,\ q=0.75$		$p=0,\ q=0.75$		$p=0,\ q=-0.75$	
		N=5	N=10	N=5	N=10	N=5	N=10	N=5	N=10
	T=5	0.6032	0.6234	0.6639	0.7551	0.5637	0.5556	0.5523	0.6432
$\{p_1, q_1\}$	T=10	0.5020	0.5134	0.6437	0.7814	0.4250	0.4579	0.4203	0.4582
	T=15	0.4905	0.5807	0.6736	0.7826	0.3350	0.3493	0.3906	0.4407
	T=5	0.5977	0.6138	0.6626	0.7596	0.5690	0.5544	0.5453	0.6517
$\{p_1, p_2, q_1, q_2\}$	T=10	0.4986	0.5143	0.6387	0.7775	0.4278	0.4592	0.4168	0.4562
	T=15	0.4924	0.5846	0.6800	0.7844	0.3438	0.3440	0.3821	0.4308
	T=5	0.5954	0.6185	0.6571	0.7475	0.5657	0.5494	0.5499	0.6424
$\{p_1, p_2, p_3, q_1, q_2, q_3\}$	T=10	0.4984	0.5175	0.6503	0.7730	0.4300	0.4536	0.4142	0.4663
	T=15	0.4939	0.5870	0.6662	0.7920	0.3425	0.3512	0.3918	0.4504

Simulation results for smaller sample size correlation of common factors $\tilde{f}_t$ and f

		$p=0,\ q=0$		$p=0,\ q=0.75$		$p=0,\ q=0.75$		$p=0,\ q=-0.75$	
		N=5	N=10	N=5	N=10	N=5	N=10	N=5	N=10
	T=5	0.6142	0.6441	0.6731	0.7821	0.5770	0.5734	0.5807	0.6689
$\{p_1, q_1\}$	T=10	0.5208	0.5376	0.6670	0.7994	0.4435	0.4785	0.4487	0.4729
	T=15	0.5267	0.5937	0.6939	0.8001	0.3552	0.3698	0.4168	0.4674
	T=5	0.6215	0.6400	0.6743	0.7812	0.5814	0.5762	0.5842	0.6649
$\{p_1, p_2, q_1, q_2\}$	T=10	0.5167	0.5354	0.6697	0.8091	0.4442	0.4779	0.4442	0.4730
	T=15	0.5268	0.6016	0.6945	0.7907	0.3628	0.3792	0.4261	0.4687
	T=5	0.6049	0.6507	0.6643	0.7816	0.5771	0.5767	0.5728	0.6625
$\{p_1, p_2, p_3, q_1, q_2, q_3\}$	T=10	0.5267	0.5435	0.6576	0.8047	0.4529	0.4717	0.4400	0.4665
	T=15	0.5294	0.5877	0.6875	0.7903	0.3469	0.3665	0.4256	0.4616

Table XIII shows estimators of correlation coefficients between common factors estimated using a new methodology $\hat{f}_t$ or the classical methodology $\tilde{f}_t$ and the true factor f_t. Correlations are calculated for panels with smaller number of columns N and observations T. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(ρ, Θ) type process.

Table XIV Simulation results for smaller sample size correlation of common components $\hat{c}_t$ and c_t

		$p=0,\ q=0$		$p=0,\ q=0.75$		$p=0,\ q=0.75$		$p=0,\ q=-0.75$	
		N=5	N=10	N=5	N=10	N=5	N=10	N=5	N=10
	T=5	0.4960	0.4922	0.5302	0.6508	0.4234	0.4372	0.4095	0.5252
$\{p_1, q_1\}$	T=10	0.3663	0.3801	0.5119	0.6734	0.2993	0.3151	0.2997	0.3107
	T=15	0.3520	0.4676	0.5664	0.6591	0.2257	0.2029	0.2882	0.3163
	T=5	0.4694	0.5115	0.5475	0.6184	0.4323	0.4196	0.4268	0.5262
$\{p_1, p_2, q_1, q_2\}$	T=10	0.3554	0.3699	0.5349	0.6434	0.2893	0.3550	0.2953	0.3541
	T=15	0.3583	0.4761	0.5450	0.6788	0.2312	0.2011	0.2534	0.2850
	T=5	0.4536	0.5078	0.5283	0.6422	0.4168	0.4200	0.4264	0.5082
$\{p_1, p_2, p_3, q_1, q_2, q_3\}$	T=10	0.3648	0.3766	0.5123	0.6384	0.3256	0.3294	0.2831	0.3399
	T=15	0.3557	0.4568	0.5333	0.6471	0.2415	0.2407	0.2871	0.3105

Simulation results for smaller sample size correlation of common components $\tilde{c}_t$ and c_t

		$p=0,\ q=0$		$p=0,\ q=0.75$		$p=0,\ q=0.75$		$p=0,\ q=-0.75$	
		N=5	N=10	N=5	N=10	N=5	N=10	N=5	N=10
	T=5	0.5137	0.5358	0.5449	0.6764	0.4622	0.4472	0.4511	0.5477
$\{p_1, q_1\}$	T=10	0.4199	0.3997	0.529	0.6737	0.3019	0.3730	0.3002	0.3690
	T=15	0.4044	0.4879	0.5475	0.6936	0.2462	0.2316	0.3122	0.3376
	T=5	0.4765	0.5351	0.5568	0.6420	0.4626	0.4715	0.4658	0.5203
$\{p_1, p_2, q_1, q_2\}$	T=10	0.4095	0.3871	0.5385	0.6983	0.3349	0.3600	0.3105	0.3577
	T=15	0.3897	0.4964	0.5922	0.6906	0.2559	0.2750	0.2986	0.3544
	T=5	0.4828	0.5252	0.5169	0.6809	0.4331	0.4533	0.4497	0.5576
$\{p_1, p_2, p_3, q_1, q_2, q_3\}$	T=10	0.4040	0.4030	0.5366	0.6884	0.3215	0.3642	0.3333	0.3440
	T=15	0.4245	0.4489	0.5601	0.6730	0.2136	0.2308	0.3186	0.3129

Table XIV shows estimators of correlation coefficients between commoncomponent estimated using a new methodology $\hat{c}_t$ or the classical methodology $\tilde{c}_t$ and the true common component c_t. Correlations are calculated for panels with smaller number of columns N and observations T. Panels are constructed using equation 6 from simulated factors, loadings and error terms; the error terms are simulated using an ARMA(ρ, Θ) type process.

Appendix B: Proofs

Henceforth, we define $\delta_{nT} \equiv \min\{\sqrt{n}, T\}$.

Proof of Lemma 1. The uniform consistency result in (8) follows directly from applying Proposition 1 in Bai (2004) to $\left\|\tilde{f}_t - H'f_t\right\| \leq \left\|\hat{F}_t - H'F_t\right\| + \left\|\hat{F}_{t-1} - H'F_{t-1}\right\|$, whence $\max_t \left\|\tilde{f}_t - H'f_t\right\| \leq 2\max_t \left\|\hat{F}_t - H'F_t\right\|$.

We now show (7). Let $e_t = [e_{1t}, ..., e_{nt}]'$ and $u_t = [u_{1t}, ..., u_{nt}]'$. Using equation (A.1) in Bai (2004, p. 164):

$$\begin{aligned}\tilde{f}_t - H'f_t &= T^{-2}\sum_{s=1}^{T}\tilde{F}_s\gamma_{n,st} + T^{-2}\sum_{s=1}^{T}\tilde{F}_s\xi_{n,st} + T^{-2}\sum_{s=1}^{T}\tilde{F}_s\left(\frac{F_s'\Lambda'u_t}{n}\right) \\ &\quad + T^{-2}\sum_{s=1}^{T}\tilde{F}_s\left(\frac{f_t'\Lambda'e_s}{n}\right) = I + II + III + IV,\end{aligned} \tag{11}$$

where $\gamma_{n,st} = E\left(\frac{u_t'e_s}{n}\right)$ and $\xi_{n,st} = \frac{u_t'e_s}{n} - \gamma_{n,st}$. As far as I, II and III are concerned, essentially the same passages as in the proof of Lemma B.2 in Bai (2004) yield $I = O_p\left(T^{-3/2}\right)$, $II = O_p\left(\frac{1}{\sqrt{nT}}\right)$ and $III = O_p\left(\frac{1}{\sqrt{n}}\right)$ respectively; the only difference is the presence of $u_t = \Delta e_t$ instead of e_t. As far as IV is concerned, $IV = \frac{1}{nT^2}\sum_{s=1}^{T}H_2'F_se_s'\Lambda f_t + \frac{1}{nT^2}\sum_{s=1}^{T}\left(\tilde{F}_s - H_2'F_s\right)e_s'\Lambda f_t = IV_a + IV_b$. Following similar passages as in Bai (2004) and exploiting the stationarity of f_t, we have $IV_a = \frac{1}{\sqrt{nT}}\left(\frac{1}{\sqrt{nT}}\sum_{i=1}^{n}\sum_{s=1}^{T}H_2'F_s\lambda_ie_{is}\right)f_t = O_p\left(\frac{1}{\sqrt{nT}}\right)$; also, $\|IV_b\| \leq \frac{1}{\sqrt{nT}}\left(T^{-1}\sum_{s=1}^{T}\left\|\tilde{F}_s - H_2'F_s\right\|^2\right)^{1/2}$ $\left(T^{-1}\sum_{s=1}^{T}\left\|\frac{e_s'\Lambda}{\sqrt{n}}\right\|^2\right)^{1/2}\|f_t\| = O_p\left(\frac{1}{\sqrt{nT}}\right)O_p\left(\delta_{nT}^{-1}\right)$; thus, $IV = O_p\left(\frac{1}{\sqrt{nT}}\right)$. The distributional result follows from noting that III is the dominating term when $\frac{n}{T^3} \to 0$; its asymptotics is studied in Bai (2004, Theorem 2).

Finally, consider (9). Using (11)

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\tilde{f}_t - H'f_t\right)u_{it} &= T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}\hat{F}_s\gamma_{n,st}u_{it} + T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}\hat{F}_s\xi_{n,st}u_{it} + T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}\hat{F}_s\left(\frac{F_s'\Lambda'u_t}{n}\right)u_{it} \\ &\quad + T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}\hat{F}_s\left(\frac{f_t'\Lambda'e_s}{n}\right)u_{it} = a + b + c + d.\end{aligned}$$

Consider $a = T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\hat{F}_s - H'F_s\right)\gamma_{n,st}u_{it} + T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}F_s\gamma_{n,st}u_{it} = a_1 + a_2$. It holds that $a_1 \leq T^{-3/2}\left(T^{-1}\sum_{s=1}^{T}\left\|\tilde{F}_s - H_2'F_s\right\|^2\right)^{1/2}\left(T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}\left|\gamma_{n,st}\right|^2\ T^{-1}\sum_{t=1}^{T}u_{it}^2\right)^{1/2} = T^{-3/2}O_p\left(\delta_{nT}^{-1}\right)O_p(1)$.

Also, $a_2 \leq T^{-3/2}T^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}\left|\gamma_{n,st}\right|\left(E\left\|\frac{F_s}{\sqrt{T}}\right\|^2\right)^{1/2}\left(Eu_{it}^2\right)^{1/2} = O_p\left(T^{-3/2}\right)$. Thus, $a = O_p\left(T^{-3/2}\right)$. Next, $b = T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\hat{F}_s - H'F_s\right)\xi_{n,st}u_{it} + T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}F_s\xi_{n,st}u_{it} = b_1 + b_2$. As far as b_1 is concerned, $b_1 \leq T^{-1}\left(T^{-1}\sum_{s=1}^{T}\left\|\tilde{F}_s - H_2'F_s\right\|^2\right)^{1/2}\left[T^{-1}\sum_{s=1}^{T}\left(T^{-1}\sum_{t=1}^{T}\xi_{n,st}u_{it}\right)^2\right]^{1/2}$; the same passages as in Bai (2003, p. 163) yield $\left[T^{-1}\sum_{s=1}^{T}\left(T^{-1}\sum_{t=1}^{T}\xi_{n,st}u_{it}\right)^2\right]^{1/2} = O_p\left(n^{-1/2}\right)$, whence $b_1 = T^{-1}O_p\left(n^{-1/2}\delta_{nT}^{-1}\right)$. As far as b_2 is concerned, the same passages as in Bai (2003, p. 163) applied to $\frac{1}{\sqrt{nT}}\frac{1}{T}\sum_{t=1}^{T}z_tu_{it}$, with $z_t = \frac{1}{\sqrt{nT}}\sum_{s=1}^{T}\sum_{k=1}^{n}F_s\xi_{n,st,k}$, yield $b_2 = O_p\left(\frac{1}{\sqrt{nT}}\right)$; thus, $b = O_p\left(\frac{1}{\sqrt{nT}}\right)$. Turning to c, $c = T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\tilde{F}_s - H_2'F_s\right)\left(\frac{F_s'\Lambda'u_t}{n}\right)u_{it} + T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}F_s\left(\frac{F_s'\Lambda'u_t}{n}\right)u_{it} = c_1 + c_2$. Using the Cauchy-Schwartz inequality and following the same passages as in Bai (2003, p. 164), $c_1 = T^{-1}O_p\left(n^{-1/2}\delta_{nT}^{-1}\right)$, and $c_2 = H'\left(\frac{1}{T^2}\sum_{s=1}^{T}F_sF_s'\right)\left(\frac{1}{nT}\sum_{k=1}^{n}\lambda_ke_{kt}u_{it}\right)$, which is $O_p\left(\frac{1}{\sqrt{nT}}\right) + O_p\left(\frac{1}{n}\right)$. Finally, $d = T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(\tilde{F}_s - H_2'F_s\right)\left(\frac{f_t'\Lambda'e_s}{n}\right)u_{it} + T^{-3}\sum_{t=1}^{T}\sum_{s=1}^{T}F_s\left(\frac{f_t'\Lambda'e_s}{n}\right)u_{it} = d_1 + d_2$. As far as d_1 is concerned, a similar logic as to c_1 yields $d_1 = T^{-1}O_p\left(n^{-1/2}\delta_{nT}^{-1}\right)$. Considering d_2, it can be rewritten as $\frac{1}{T}H'\left(\frac{1}{T}\sum_{s=1}^{T}F_se_s'\right)\left(\frac{1}{nT}\sum_{k=1}^{n}\lambda_kf_{kt}u_{it}\right)$, which is $O_p\left(\frac{1}{\sqrt{nT^3}}\right) + O_p\left(\frac{1}{nT}\right)$. Putting all together, (9) follows.

Proof of Proposition 1. By definition, $\tilde{\lambda}_i - H^{-1}\lambda_i = \left(\sum_{t=1}^{T}\tilde{f}_t\tilde{f}_t'\right)^{-1} \times \left[\sum_{t=1}^{T}H'f_tu_{it} + \sum_{t=1}^{T}\tilde{f}_t'\left(\tilde{f}_t - H'f_t\right)\lambda_i + \sum_{t=1}^{T}\left(\tilde{f}_t - H'f_t\right)u_{it}\right] = \left(\sum_{t=1}^{T}\tilde{f}_t\tilde{f}_t'\right)^{-1}(I + II + III)$. Consider the denominator. Lemma A.1 in Trapani (2012a) states that $\sum_{t=1}^{T}\left\|\tilde{f}_t - H'f_t\right\|^2 = O_p\left(T\delta_{nT}^{-2}\right)$ and $\sum_{t=1}^{T}\left(\tilde{f}_t - H'f_t\right)'f_t = O_p\left(\sqrt{T}\delta_{nT}^{-1}\right) + O_p\left(\frac{\sqrt{T}}{n}\right)$. Hence, $\sum_{t=1}^{T}\tilde{f}_t\tilde{f}_t' = H'\sum_{t=1}^{T}f_tf_t'H + o_p(T) = O_p(T)$. As far as the numerator is concerned, $I = O_p\left(\sqrt{T}\right)$ by a CLT. Using the same arguments as for the denominator, $II = O_p\left(\sqrt{T}\delta_{nT}^{-1}\right) + O_p\left(\frac{\sqrt{T}}{n}\right)$. This entails that $\tilde{\lambda}_i - H^{-1}\lambda_i = O_p\left(T^{-1/2}\right) + O_p\left(n^{-1}\right)$. Finally, $III = O_p\left(n^{-1/2}\right) + O_p\left(T^{-3/2}\right)$ using (9). Putting all together, the rate of convergence follows; the limiting distribution follows from noting that, when $\frac{\sqrt{T}}{n} \to 0$, the dominating $O_p\left(T^{-1/2}\right)$ term is $\left(H'\sum_{t=1}^{T}f_tf_t'H\right)^{-1}\left(\sum_{t=1}^{T}H'f_tu_{it}\right)$.

Proof of Theorem 1. In the proof, we omit H for simplicity when this does not cause ambiguity. We start by showing that $\left\|\hat{\Sigma}_F - H'\Sigma_FH\right\| = O_p\left(\frac{h}{\sqrt{T}}\right) + O_p\left(\frac{h}{n}\right) + O_p\left(\frac{1}{h}\right)$. Note first that, by definition,

$\Sigma_F = \gamma_0^F + \sum_{j=1}^{\infty} \left(\gamma_j^F + \gamma_j^{F\prime}\right)$, whence

$$\begin{aligned}\hat{\Sigma}_F - \Sigma_F &= \left(\hat{\gamma}_0^F - \gamma_0^F\right) + \sum_{j=1}^{h} \left(1 - \frac{j}{h+1}\right) \left[\left(\hat{\gamma}_j^F + \hat{\gamma}_j^{F\prime}\right) - \left(\gamma_j^F + \gamma_j^{F\prime}\right)\right] \\ &\quad - \sum_{j=1}^{h} \left(\frac{j}{h+1}\right) \left(\gamma_j^F + \gamma_j^{F\prime}\right) - \sum_{j=h+1}^{\infty} \left(\gamma_j^F + \gamma_j^{F\prime}\right) \\ &= I - II - III.\end{aligned}$$

Consider I. Focusing on $\hat{\gamma}_0^F - \gamma_0^F$, we have $\hat{\gamma}_0^F - \gamma_0^F = T^{-1} \sum_{t=j+1}^{T} \tilde{f}_t \tilde{f}_t' - \gamma_0^F = \left(T^{-1} \sum_{t=j+1}^{T} f_t f_t' - \gamma_0^F\right)$ $- T^{-1} \sum_{t=j+1}^{T} \left(\tilde{f}_t - f_t\right) f_t' - T^{-1} \sum_{t=j+1}^{T} f_t \left(\tilde{f}_t - f_t\right)' + T^{-1} \sum_{t=j+1}^{T} \left(\tilde{f}_t - f_t\right) \left(\tilde{f}_t - f_t\right)' = I_a + I_b +$ $I_b' + I_c$. The CLT yields $I_a = O_p\left(T^{-1/2}\right)$; as far as I_b and I_c are concerned, Lemma A.1 in Trapani (2012a) entails that they are both $O_p\left(n^{-1}\right) + O_p\left(T^{-2}\right)$. The same holds for $\hat{\gamma}_j^F - \gamma_j^F$, so that putting all together $I = O_p\left(hT^{-1/2}\right) + O_p\left(hn^{-1}\right)$. Standard arguments yield $II = O\left(h^{-1}\right)$ and $III = o\left(h^{-s}\right)$. The Theorem follows from $\hat{\lambda}_i - H^{-1}\lambda_i = O_p\left(T^{-1}\right)$.

This page is intentionally left blank.

This page is intentionally left blank.

4 Large Dimentional Panel Interpolation Using EM algorithm

Ekaterina Ipatova Lorenzo Trapani

Cass Business School, City University London

April 9, 2014

Abstract

The presense of structural and non-structural omitted observations is an issue that features regularly in large dimensional panels literature. There are a wide range of methodologies that have been proposed in recent times, however a majority are developed around treating the omitted observation problem on case-specific basis. This involves extracting time-series from the panel in order to fill in the missing observations and in effect, disregarding the presence of cross-sectional correlation between the series. This approach could potentially distort the results of any factor based analysis.

Our study proposes substituting the omitted observations to the panel with respect to the common trends between the variables. Our methodology could also accommodate any type and proportions of omitted observations. Given a large number of variables in the factor models we believe that our methodology enables the preservation of richer dynamics. We carry out a finite sample analysis with competing strategies and finally provide an empirical study to illustrate the superiority of our models.

4.1 Introduction

In this paper, we develop an approach that is able to remedy a problem of omitted observations in the large dimensional panel irrespectively of the number and location of the gaps. Until recently academic publications paid little attention to the problem of missing observations, although the existence of the problem was widely acknowledged in the financial industry. The problem was commonly solved using simple data manipulations, for example, listwise deletion, which deletes observations on the specific date for all variables (used in 94% of papers before 2000), substitution of the variable mean (King (2013)), or even best guess imputation. More advanced approaches normally included modelling of bridge equations, which apply a linear regression between fragmented variable and corresponding counterparts; the fitted values were used on the place of omitted variables. Also modern techniques include state-space models or MIDAS approach (Foroni and Marcellino (2013)). All models, although popular are applied to single variable with irregular frequency. Multivariate alterations of the techniques often combine factor approach with one of methodologies. However once again the multivariate approach means application of interpolation to single variable using high-frequency multivariate panel.

Current research offers solutions to the problem of missing observations across multivariate panels of variables with cross-sectional and time-series dependence. The methodology applicable to three patterns of missing observations within a panel. First, mixed frequency pattern, when some variables become available with higher a frequency than others. For example, US GDP is only available quarterly while US energy consumption is published on a monthly basis. Second, ragged-edge pattern (see Wallis, 1986), when missing observations are unobserved for a few most resent observations all across the panel. Third, random missing observations: when some of the variables are simply unpublished on certain dates. This is particularly noticeable when we work with developing countries where the historical data is collected irregularly.

A procedure builds on bridge equations, expected maximisation and more importantly on the common trend methodology, discussed in the theory of factor models. The idea is to build a bridge equation between a common factor from a mixed frequency panel and a factor from counterpart panel; the fitted values are optimised using expected maximisation approach and are used to fill missing observations across a panel. This methodology helps to achieve backcasting, nowcasting and forecasting of variables in the panel; therefore, the approach is fairly general and helps to solve a majority of the issues related to unbalanced panels. The idea is not unique, and it builds on the earliest work on unbalanced panels by Stone (1947), and seminal research by Stock and Watson (2002), amongst others. Most of the studies

explore only the third type of missing observations pattern, when only a few observations are missing at random. Current research proposes methodology that aims to provide information to the entire panel disregarding the proportion or pattern of missing observation.

Relevance of the topic is discussed during detailed critical evaluation of the literature in the next section. Next we provide a description of the theorem with proofs that allow the existence of the technique. A more detailed description of the methodology is provided in the latter part of the paper. Section 4.4 presents the results for the simulation exercises as well as a discussion of the results of some practical application to energy markets.

Energy markets are chosen as they necessitate a large number of parameters to determine fundamental forces behind commodity prices. This is as a traditional pricing approach advocates assessment of current supply and demand of energy commodity; if supply is far greater then demand prices tend to decline. Supply and demand is determined by a vast number of factors related to the global economy and the subtelities of the geopolitics which are hard to quantify. In order to have a general model based on fundamental determinants, we need to evaluate the market movements using a large data set. Modern databases offer a wide range of economic, financial and geopolitical data that provide valuable information for this analysis; however it is not uncommon that it misses a large proportion of observations. Our proposed methodology is well suited to fill missing observations across large dimensional and cross-correlated panel such as mixed frequency energy panel, which makes it optimal illustration for the approach.

The results can be summarised as follows. This paper contributes to the literature on interpolation techniques of large dimensional panels. To the best of my knowledge the methodology proposed by Stock and Watson (2000) and alternative methodology by Forni et al.(2001) are the only ones that use interpolation of large dimensional panels. The research attempts to extend Stock and Watson's (2000) technique.

The research contributes to both theoretical literature, by providing a new theorem and methodology, and empirical literature by establishing new application for factor interpolation. Our most interesting finding is the newly formulated methodology, which can help to remedy a problem of mixed frequency data, ragged edge data and random omitted variables.

4.2 Literature review

A substantial theoretical and empirical pool of literature identifies a number of widely applied methodologies which solve a problem of missing observations. I present a critical evaluation of available techniques and identify gaps in the literature.

The majority of the techniques available for the problem of omitted observations focus on interpolation of single variables (see, survey by Foroni and Marcellino (2013)). To the best of my knowledge early models were available exclusively to single variables; among them are splines and aggregation. These models are able to provide a quick solution, however they also create a range of problems as they rely on the assumption that information in the high-frequency variable is reflected in the low-frequency representation. This assumption largely depends on the underlying variable flow, such that only variables with low volatility can satisfy it, otherwise the risks of losing important information run high.

Alternatively, one could use bridge equation methodology, which the relaxed assumption above. Bridge equation methodology is still one of the most used for short term forecasting (see for example, Baggi, Golinelli, Parigi (2004) and Diron (2008)), however they are also effective in addressing the problem of mixed observations. The examples of application mostly include forecasting of economic data, and in particular GDP; study by Trehahn and Ingenito (1996) used bridge equations to forecast current US GDP; Baffigi et al. (2004) apply bridge equations to GDP of Euro Area. The critics point out that bridge equations are essentially statistical models, which are prone to theoretical misspecification (see Foroni and Marcellino (2013)). In general, bridge models work well forecasting of the single variable, forecasting of large dimensional panels with ragged-edge is too hard to handle, as every variable has to be forecasted separately.

More sophisticated techniques for mixed frequency data include MIDAS models and modification of state-space models. Two methodologies are often compared, for example, studies by Kuzin, Marcellino and Schumaher (2011), Ghysels et al. (2005), Ghysels, Santa-Clara and Valkanov (2006)); they conclude that state space MF-VAR (Mariano and Murasawa (2010)) model tend to do better on long-term horizon, while MIDAS perform better for shorter periods. The main distinction between the two approaches is that MIDAS tend to deal with single dependent variable, while MF-VAR can model endogenaity between multiple variable similar to classical VAR. the main drawback of the state-space models, including Kalman filter (see Mittnik and Zadrozny (2005)), is that state-space models can be overly parameterised and need to estimate large number of parameters, which leads to the degrees of freedom problem. MIDAS is simpler,

however it cannot model interdependency between models.

From the evidence above it is obvious that the majority of the models are focused on solving a problem of missing observations for one or only a few variables, and for a few of the most recent observations (short term forecasting horizon). Therefore, there is a need for the approach that can be applied for multivariate datasets. Majority of the models above have multivariate representation, combining factor models with the methodology. Among others, there is a paper by Doz et al. (2011) on the application of bridge equations with factor trends, state-space VAR models with factors (see, Banbura and Rünstler (2011)), and factor MIDAS (see, Marcellino and Schumacher (2010), Altissimo et al. (2010)).

The techniques above rely on balanced high-frequency panels which are used for interpolation of one single variable. There are only a handful of techniques that can help to minimise the impact of a problem of missing observations inside the panel used for factor model application. Among most used are techniques described by Stock and Watson (1999), Altissimo et al (2001)); however they are applicable to panels with low overall proportion of missing observations. The technique that can remedy a larger proportion of missing variables (more than 75%) has not been identified during my literature review. Therefore, current research aiming to provide theoretical and empirical evidence requires consideration of a potential remedy for highly unbalanced panels. The methodology is based on the factor model literature that is discussed in detail in chapter 2. Here I present an extension to the factor interpolation techniques of the modern literature.

The importance of the topic is evident as many leading studies provide strong evidence that implementation of a factor structure results in smaller MSE than competing techniques which are based on simple bridge models or structural models. Among others we refer to Stock and Watson (1998, 2002b) on diffusion index interpolation, Forni et al. (2000) and Armah and Swanson (2010, 2011). Various practitioners such as the Federal Reserve of Chicago, the US Treasury, the European Central Bank, the European Commission, and the Centre for Economic Policy Research all acknowledge the importance of a factor (or diffusion index) methodology in their models. Armah and Swanson (2010, 2011) and Stock and Watson (2002a, 2006) evaluate the utility of factor models and diffusion indexes for nowcasting and problems of "ragged-edge" data.

Recent surveys (see Boivin and Ng (2005), Eickmeier and Ziegler (2008), Marcellino and Schumacher (2008)) indicate that the diffusion index framework generally consists of two stages. The first step involves estimation of the model. Commonly referred to as "step (E)", it and involves estimation of the

latent factors from the large dimensional panel. We optimize factors by applying Expected Maximization or Maximum Likelihood methods. These minimize the error between estimated and true data points. The second step involves multiplication of latent factors on factor loadings which fill previously omitted observations. For simplicity, we refer to the second stage as "step (F)", filling low frequency series with missing observations. Bovin and Ng (2005) evaluate the sensitivity of step (E) with the methods used for factor estimation and find that none of the existing methods produce superior results. The results indicate that any currently available method for factor extraction should provide solid results; thus the choice of approach should not affect the final output.

Current literature recognises two main methodological branches that use the factor models method to interpolate unbalanced panels. The first approach builds on the work by Stock and Watson (1998,2002b) and uses factor estimation procedure that involves the application of the principal component analyses to extract latent factors. Principal component and factor analyses were recognised as a standard evaluation tool for large dimensional data from the early stages of econometrics development (see Anderson (1958)). More recently, we saw an increasing number of academic papers aiming to develop additional methods for analyses of large dimensional data by the application of PC. Most striking contributions were made by Bai (2003,2004), who developed an asymptotic theory for estimated factors, loadings and common components. They define properties of distributions for all components and more importantly relax the assumptions of the model for the large dimensional panels. This finding is crucial for the purpose of current research given that proofs of the research were developed on the basis of original assumptions in Bai (2003). Bai and Ng (2002, 2006, 2011) examine topics related to the optimal factor model structure and establish techniques to identify optimal number of factors, whereas later authors worked on defining confidence intervals for the components.

Stock and Watson's framework (henceforth SW) is widely used by academics and practitioners. For example, the paper by Bernanke and Boivin (2003) uses SW methodology to implement factor interpolation for real-time GDP. The procedure reflects the complexity of the problem related to establishing monetary policy in a situation when final GDP figures are published with considerable delay. SW frameworks apply to various models currently used by practitioners. Amongst others, we found static factor interpolation in the methodology of the Federal Reserve Bank of Chicago's Activity Index (CFNAI) and the US Treasury model developed and published by Kitchen and Monaco (2003).

Factors extracted by PC are interpolated using random values and later commonly improved using the Expected Maximization procedure. EM aims to minimize the error between true and interpolated

values in the factors. Application of EM is widely used for interpolation and extrapolation in the models where frequency is mixed (see for example Mitchell et al. (2005)). It was originally developed in the work by Friedman (1962), Chow and Lin (1971) where authors apply EM procedure to minimise the error between the estimated value for missing observation and true value in the series. EM technique in factor model framework is applied directly to the factors. Schumacher and Breitung (2008) employed a method of factor EM interpolation where latent factors were first extracted and later interpolated using random draws from a standard normal distribution as "a first guess" for EM initialization. Alternatively, Biernacki et al (2003) described a procedure of random initialization based on previous runs of EM that are only available for variables without any data irregularities. Angelini et al. (2006) describe the development of EM interpolation with respect to factor models; Marcellino et al. (2007) provide a comparison to other approaches.

Alternative methodologies for latent factor extraction are developed in the second branch of the literature. The procedure is appealing to empirical research since it deals with data irregularities, non-synchronicity and publication lags of the data (primarily GDP) in the Euro Area. It is based on the work by Forni et al. (2000, 2001), examining generalised dynamic factor models which are similar to Stock and Watson (1998) diffusion index models. Forni et al. (2000, 2001) work applied inverse Fourier transformation to acquire common components instead of spectrum decomposition. More importantly, the procedure differs from the SW owing to the fact that it defines common components in the frequency domain, allowing the use of fluctuations with waves with periods of over a year and therefore, filtering out short term noise. As a result, frequency models allow the recognition of established long term cyclicality of the common components which then become a template for missing observations.

The cyclicality of common components is not the only parametric specification that can be imposed on factors. Recent developments in the academic literature discuss a possibility of ARMA process pre-specification of factors (see Mariano and Murasawa (2010)). In addition, Camacho et al (2012) additionally examine the possibility of applying Markov switching models to dynamic factors.

As we stated above, this factor methodology is unique as it attempts to remedy omitted observations over cross-sectional dependent panels. Concurrently, this technique is predominatly applied to the panels with low proportion of missing observations. In our research we attempt to improve this condition by suggesting methodology of non-random initialisation of expected maximisation approach. This methodology should help to improve the precision of optimisation procedures and allow to interpolate panels with a large portion of mixed–frequency panels and for short term forecasting of ragged-edge panels.

Based on the findings of Bovin and Ng (2005) we feel that there will be no qualitative difference in the methodology of our study if we use any of the popularly available techniques for common components estimation. We therefore follow the SW system and estimate factors using a PC methodology. In the following chapter we describe a methodology that uses factor values to fill missing observations in the related panel. Initial values are then improved using the EM approach. Next, we present a number of Monte Carlo simulations that aim to match factor EM techniques with available competing methodologies. We also perform an empirical application to a dataset from the energy markets. We carry out now-casting of prices for oil refinery products using large dimensional panels consisting of a wide variety of data representing oil supply and demand determinants along with other relevant macroeconomic data series.

4.3 Methodology

In this section, we discuss the methodology of the Factor-EM procedure. We begin by introducing a common form of a factor model that follows the assumptions from Bai (2003,2004). Next we introduce data irregularities and propose a methodology for their identification. We offer a detailed study of the unique initialisation method developed specifically for this approach, including rigorous proofs. Finally, we discuss all stages of the Expected Maximisation procedure.

4.3.1 Generalized equidistant factor model

We define the generalized factor model in a manner that is consistent with the formulation and notations of Stock and Watson's (2002b) classical framework. We define a matrix X which is a large dimensional panel of N columns of constituent vectors over T time periods. The dimension T represents equidistant time intervals indicative of a balanced matrix X and neglects any data irregularities. Decomposition of the matrix X results in two features, the common component denoted by (ΛF) and the second, the idiosyncratic factor (E). Common component(s) describe common trends amongst vectors with the idiosyncratic factors corresponding to individual fluctuations within the series. The common components are linear combinations of the column vectors N and are, therefore, devoid of any economic interpretations. A factor model has the following representation;

$$X = \Lambda' F + E \tag{12}$$

$$F_t = \Phi_p F_{t-p} + u_t \tag{13}$$

Common components feature common factors F and factor loadings Λ. The matrix of the common factors F is of the dimension $(T \times k)$ and contains common trends between column vectors; matrix Λ $(N \times k)$ provides a measure of association of each column vector to the common trends; Bai and Ng (2002) methodology determines the optimal number of trends k. The idiosyncratic component is a matrix E size $(T \times N)$. We consider two options: first for stationary panels, such that coefficient p is low and reflects weak time-series dependence and second, non-stationary panels, when coefficient p may be large (more than 3) reflecting strong autocorrelation in the factor trend.

Model 12 follows assumptions published in Bai (2003, 2004). These assumptions are more relaxed in comparison to the earlier factor models, which assumed that the error terms were iid. Generalised factor models include realistic allowances for idiosyncratic components to satisfy market conditions. In order to be consistent with the literature, we follow a list of assumptions from Bai (2003, 2004). Section 2.3.1 outlines a full list of assumptions; Appendix A section 6 lists additional assumptions for asymptotic proofs.

We differentiate assumptions for stationary and non-stationary panels. Non-stationary panels include column vectors specified by $I(1)$ autoregressive process of order one (*assumptions A-D*, section 2.3.1); stationary panels include $I(0)$ vectors (*assumptions E-H*, section 2.3.1). *Assumption A* is responsible for the properties of the autoregressive process, creating a dynamic between time dimensions of common factors. This is a vital modification in comparison to the earlier models, which assumed independence across time dimension of common trends. It does not, however, relax the relationship between common factors and the panel, which is perceived to be static.

Assumption B sets the properties of factor loadings to ensure that each column vector has a unique representation. The matrix of factor loadings can be random providing that two conditions hold: it is independent of common factors and error terms and the existance of the fourth moment.

Assumption C assures that model 12 reflects the observable characteristics of the data on the market. They relax conditions for idiosyncratic errors, which now can have weak serial and cross-sectional correla-

tion, and heteroskedasticity. It is a substantial improvement in comparison to the iid errors of basic factor models, or to the approximate factor model assumption that singularly allows for weak cross-sectional correlation in the error terms. *Assumption D* establishes independence between factor loadings and error terms.

Assumptions E to H define components in the stationary panels. *Assumption E* defines bounds for the dynamics across time dimension in common trends. *Assumptions F and G* are identical to B and C; finally *Assumption H* allows weak dependence between errors and common components in stationary panels.

Overall, the assumptions of the generalised factor model allows estimation of consistent parameters given the properties of financial market data. Early models do not allow for vital characteristics of market data, such as serial – correlation of idiosyncratic factors, and therefore, are more restricted in their applications. Many factor-interpolation procedures (e.g. Schumacher and Breitung (2008)) follow the earlier modelling approach and therefore the outputs from these models need to be considered in lieu of the implied assumptions.

To estimate common components and idiosyncratic errors for model 12 the methodology employs principal components (hereafter PC) method. In removing scale effects, the time series is demeaned. Following PC methodology we use spectrum analysis on the variance-covariance matrix of the large dimensional panel X. In doing so, we attempt to decompose the covariance matrix into their constituent eigenvalues and eigenvectors. The factors are $\sqrt{T}$ times eigenvectors corresponding to the eigenvalues of the $T \times T$ matrix XX' for stationary panels; it is T times eigenvectors corresponding to the eigenvalues of the $T \times T$ matrix XX' for non-stationary panels. Factor loadings estimated using equation $\Lambda = FX/T$ and common components are $F\Lambda'$.

For long panels, optimal estimation method calculates the matrix of factor loadings Λ which is $\sqrt{N}$ times eigenvalues corresponding to the eigenvectors of the $N \times N$ matrix, then determines common factors according to the formula $F = X\Lambda/N$. Common components are identical to $F\Lambda'$. The aim of the method is to minimise idiosyncratic components and simultaneously keep the number of common components to a minimum.

4.3.2 Mapping data irregularities

In order to ensure consistency of the model parameters, we have previously defined the generalized factor model along with a list of relevant assumptions. In this section, we introduce data irregularities within the large dimensional panels and propose our methodology to mitigate this issue.

Large dimensional panel can be subject to the problem of data irregularity due to the data gathering and reporting problems. The complexity of the data collection process regularly leads to the "publication lags", "low" frequency of the data or data unavailability. As a result, panels become unbalanced; and the pattern of missing observations has three potential structures: first an unsystematic structure when observations in the column vectors are unavailable in a random order. It is associated with rare events taking place due to an anomaly in data gathering. More severe cases have systematic structure of data unavailability such as either "low frequency reporting" or "publication lags". Low frequency reporting can be identified only in the content of the ideal frequency of the panel time dimension. It occurs when data for the single column vector is available at a lower frequency than the rest of the vectors. A common example that illustrates the issue with mixed frequency usually involves GDP which is only available at a quarterly frequency. Additionally, a large panel could also involve a "ragged-edge" problem, when data has few missing observations at the beginning or the end of every variable in the panel. In other words, it is a ragged structure of the panel bottom (top) border as a result of unavailability of the data after (before) a certain date. The recent observations are unavailable possibly due to publication lags. Additionally, the older data may be unavailable due to issues relating to data gathering and verification over the entire period, such as in times of political uncertainty in emerging markets.

Our methodology is an unified approach that addresses all three types of data irregularities. We start formalisation of the approach by introducing some data irregularities to the panel X. We define panel X^* size $(T^* \times N)$ such that it includes some missing observations, where the type of data irregularities are not relevant. The goal is to convert panel X^* that includes missing observations to the balanced panel X of the size $(T \times N)$. In order to do that we introduce selection matrix A, such that the following relationship holds:

$$X = A'X^* \tag{14}$$

Selection matrix A size $(T^* \times T)$ is known and is required to have a full rank. Without data irreg-

ularities matrix A is an identity matrix. To clarify the methodology we consider an example that aims to to interpolate three quarters of the UK and the US GDP to nine monthly observations (three months per quarter). Following the methodology we formalise the problem by considering a panel of data X^* size (3×2) that we transfer to panel X size (9×2). We apply equation 14 using the selection matrix resulting in panel $X^{(0)}$ below:

$$A'X^* = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}' \times \begin{pmatrix} UK_{1,1} & US_{2,1} \\ UK_{1,2} & US_{2,2} \\ UK_{1,3} & US_{2,3} \end{pmatrix}$$

$$X^{(0)\prime} = \begin{pmatrix} UK_{1,1} & 0 & 0 & UK_{1,2} & 0 & 0 & UK_{1,3} & 0 & 0 \\ US_{2,1} & 0 & 0 & US_{2,2} & 0 & 0 & US_{2,3} & 0 & 0 \end{pmatrix}$$

Matrix X^* expands using selection matrix A and forms preliminary matrix $X^{(0)}$ same size as X. Observations of the preliminary matrix equals zero in those instances of missing observations. Selection matrix A is always known; however, it does not have a unified form. Every time we perform an interpolation procedure for the new panel we have to study X^* and modify the selection matrix such that it fits appropriate matrix $X^{(0)}$. For every interpolation, matrix $X^{(0)}$ should position available observations at their expected locations and supply zeros for missing observations.

To achieve a balanced panel X we have to modify zero values across panel $X^{(0)}$. Initial transformation substitutes zero values to the approximation of the missing observations, thus replacing panel $X^{(0)}$ to $X^{(1)}$; next the Expected Maximization procedure optimizes initial approximations of missing values in panel $X^{(1)}$. Before we describe detailed methodology of the EM process, we would like to draw attention to the transformation of $X^{(0)}$ to $X^{(1)}$.

4.3.3 Factor-Initialisation

The initialisation (transformation) procedure has a few variations in the literature. Biernacki et al (2003) conducts a survey of random initialisation procedures (henceforth RI) in which he describes the methodology and evaluates perturbation of variates, including classification of the EM algorithm, a Stochastic EM algorithm and short runs of EM, among others. RI methodology fills zero values of

the panel X_0 with random draws from standard normal distributions. Later the EM algorithm is used for optimization. Our study proposes a novel methodology for initialization of the EM procedure that provides non-random initialisation (henceforth NRI) values which have greater proximity to the true values than RI.

The structure of the generalised factor model is in the core of NRI methodology. Consider re-defining panels X and X^* according to the structure of generalised factor model, equation 12. We argue that the difference between two panels originates in the common trends. Factor loading should remain relatively similar, such that they are responsible for the unique representation of column vectors that are equivalent for both panels. In other words the following relationship holds:

$$X^* = \Lambda^{*'} F^* + E^* \tag{15}$$

$$X = \Lambda^{*'} F + E \tag{16}$$

Where X^* and F^* are panels and common factors in unbalanced panels, X and F correspond to the balanced panel. To the best of our knowledge, the assumption that factor loading is constant does not have a significant impact on the initialisation procedure. The varying factor loading can therefore be easily achieved. For the purpose of this paper, we fixed factor loading such that it is identical in balanced and unbalanced panels. We consider variable factor loading to be a subject for further extension to current research.

To apply Factor-NRI we introduce panel Y that satisfies two conditions: first that it is a balanced panel of size $(T \times N_Y)$ or in other words it has exactly the same length as panel X; second panel Y has a high correlation with the panel X. For example, panel X can consist of medium term heating oil futures available only after 1995 while panel Y consists of short term futures available from 1990. This is an extreme example. The methodology does not require such a proximity for applications since the aim of the procedure is to provide initial values that will be optimised. Below we define panel Y.

$$Y = \Psi^{*\prime} G + \Sigma \tag{17}$$

$$Y^* = \Psi^{*\prime} G^* + \Sigma^* \tag{18}$$

Where Ψ^* corresponds to factor loading, Y is the observed balanced panel size $(T \times N_Y)$ and G is a matrix of common factor $(T \times k)$; k for panel Y is equal to panel X; Σ is the idiosyncratic factor. We use selection matrix A which maps missing observations in panel X^*. We apply selection matrix A to balance panel Y using equation 14. As a result we obtain panel Y^*, and following the transformation has an identical pattern of missing observations as X^*. We apply PC analyses to Y^*, to find G^*.

The identification of a suitable panel Y is a key for successful application of the technique. The identification of the panel Y varies depending on the empirical characteristics of the data. The application of the NRI techniques to term structure data (similar to the heating oil futures) results in the direct identification of Y, as part of the term structure variables which do not have missing observations. For the remaining cases, we suggest close analyses of the dataset and separation of the original dataset on variables with omitted observations (panel X) and without (panel Y). Theoretically, justification for this operation is that all variables in factor panel data should share a common trend, otherwise identification of the few common trends in the panel is impossible. Therefore, the commonality of the trends is a common characteristic between variables with missing observations and without given that they are a part of one complete factor panel dataset. We have to point out that the variables in factor panels are selected from the basis of the theoretical model, in that they are not random variables, and are selected for an analyses of the particular event. Together they constitute our justification for the assumption that the original dataset can be split on the panel X and Y, such that they will satisfy the assumptions. This definition of the panel Y should allow for easy identification and application in the method. The chapter on empirical application will demonstrate this approach in action. In comparison to the competing approaches, NRI does not induce additional problems associated with the identification of panel Y, providing that the original dataset can be separated on panels X and Y. In practice the majority of large dimensional datasets have complete variables and variables with missing observations, which permit the proposed separation. NRI method is applicable to both stationary and non-stationary panels X and Y; however, we avaluate NRI application separately in two cases.

The first option considers panel X consisting of non-stationary variables.

Following equation 15 and 18 we define the relationship between two estimators such that:

$$\hat{F}^* = \hat{\rho}^* \hat{G}^* + \Upsilon^* \tag{19}$$

where $\Upsilon^* \sim I(0)$ meaning that $\hat{F}^*$ and $\hat{G}^*$ are cointegrated; if $\Upsilon^* \sim I(1)$ we have to consider first-differencing or de-trending all column vectors in the panels X and Y, leading to both X and Y being stationary panels. Factor-NRI theory for stationary panels is described further. $\hat{F}^*$ is an estimated matrix $(k \times T^*)$ of common factors extracted from panel X^*; $\hat{G}^*$ is a matrix $(k \times T^*$) of common factors extracted from panel Y^*; $\hat{\rho}^*$ is a vector $(k \times 1)$ that represents correlation coefficients between $\hat{F}^*$ and $\hat{G}^*$.

Theorem 2 *Given assumptions A-G from Bai(2004), for* $1/\sqrt{N} \to 0$ *limiting distribution of correlation coefficient* $\hat{\rho}^*$ *given by:*

$$T(\hat{\rho}^* - \rho^*) \xrightarrow{d} \sigma_e[H_G \int_0^1 B_G B_G' H_G^{'}]^{1/2} N(0,1)$$

where B_G *is vector of Brownian motions defined in Bai(2004), Assumption A2 and* σ_G *defined in Appendix B;* $H_G = \left(\frac{\hat{G}'\hat{G}_u}{T^2}\right)\left(\frac{\Lambda'\Lambda}{n}\right)$

Limiting distribution implies that asymptotic normality generally holds. In practice researchers should feel comfortable using a mixed-normal distribution for approximation of the correlation coefficients. The limiting distribution is true for large dimensional panels. Based on the lemma 2:

Proposition 2 *As* $(N,T) \to \infty$, *let assumptions A-G Bai(2004) hold*

$$T(\hat{\rho}^* - \rho^*) = O_p(1) + O_p(1/\sqrt{N})$$

Providing that G^* and F^* are observable, ρ^* can be estimated with the rate of convergence $\sqrt{N}$ by the least squre method (according to proposition 2). Based on 17 and 19 we approxiate F :

$$\hat{F} = \hat{\rho}^* \hat{G} + E \tag{20}$$

Next we use equation 16 to creat panel $X^{(1)}$.

The second option considers panel X consisting of stationary variables. For the stationary option we use the same methodology as described in the first part of section 4.3.3 to arive to the results in 20. The difference is due to a variation in asymptotic results. The limiting distribution for stationary case is described in Theorem 3:

Theorem 3 *Given assumptions A-G Bai(2003) and for* $(N,T) \to \infty$

$$\sqrt{T}(\hat{\rho}^* - \rho^*) \xrightarrow{d} \Sigma_G^{-1}\Phi_i^{1/2}N(0,1)$$

where Φ_i *is defined in Bai(2003) p144,* Σ_G *defined in Bai(2003) p.141, which are presented in Chapter 6 of the thesis*

Providing that $(N,T) \to \infty$, then theorem 3 and the rate of convergence will be derived in appendix A. The rate of convergence is summarised in proposition 3.

Proposition 3 *As* $(N,T) \to \infty$*, it holds that*

$$\sqrt{T}(\hat{\rho}^* - \rho^*) = O_p(1) + O_p(\tfrac{\sqrt{T}}{N})$$

Assumptions for ρ coefficient After careful investigation of the coefficient ρ, we construct a set of reasonable assumptions that are then summarized within the list of *Assumptions K*:

Assumption K:

(i) $\rho^* = \rho$, where ρ is a correlation coefficient between $\hat{F}$ and $\hat{G}$

(ii) ρ is stable over period T

(iii) $\lambda_i^* = \lambda_i$, where λ_i is a set of factor loading for panel X_{it} and λ_i^* are factor loadings for X_{it}^*

Assumption K (i) relates to the conditional correlation coefficient between the two common trends and states that ρ has to be stable between different frequencies. The assumption may seem strong and therefore we feel that in future enhancements to our study, a framework that would allow for Dynamic Conditional Correlation (see Engle(2002)) would mitigate this assumption and improve performance. *Assumption K (ii)* offers stability to the coefficient ρ over time; this can be determined by using one of the structural stability tests. *Assumption K(iii)* ensures that factor loadings are constant for low and higher frequency models. Factor loadings can be modeled in a similar way to common factors by using a correlation coefficient between factor loading from panels X and Y, thus *assumption K(iii)* can be relaxed.

4.3.4 Factor EM-algorithm

To perform the Expected Maximization procedure we have to obtain panels X^* and $X^{(1)}$, as well as the common factor F and factor loading Λ and panel-specific selection matrix A. Next, for the j^{th} iteration of i^{th} variable we update values for $X^{(1)}$, such that $X^{(1)}$ becomes X^j after every j^{th} iteration. The procedure follows equation 21:

$$X^{(j)} = \hat{F}^{(j-1)}\hat{\lambda}^{(j-1)} + A'(AA')^{-1}(X^* - A\hat{F}^{(j-1)}\hat{\lambda}^{(j-1)}) \tag{21}$$

We consider equation to be the E-step (expectation step) of the procedure. E-step follows Stock and Watson's (2002a, p. 156) approach. To initialise step-E we utilize the initial matrix $X^{(1)}$ and common factor F which are approximated using Factor-Initialisation procedure.

The principal components method is applied to panel $X^{(j)}$ such that new estimates for common factor F and factor loading Λ are obtained. The new common factor and factor loading are returned to step-E and the procedure is repeated until they converge. The procedure is repeated until the maximum percentage change of the variables' estimates from step j and $j+1$ are larger than 10^{-5}.

4.4 Results

4.4.1 Simulation

The finite sample properties of our methodology are assessed via a Monte-Carlo simulaiton. The simulation starts from data generating process $F_t = F_{t-1} + u_t$ (where u_t are iid $N(0,1)$), creating the single common factor F ($r = 1$); this is non-stationary and does not have omitted observations. Common factor G for panel Y is estimated using pre-determined coefficient ρ; the equation of common factor G is $G = \rho * F$. Factor loading Λ and Ψ in panels X and Y are determined using random drowns from a uniform distribution. Error components of both panels are generated in a similar way to the ARMA process and thus allows aweak time-series correlation. Non-stationary panels X and Y are obtained by generating common factors, loadings and error terms using equations 16 and 17:

$$X = \Lambda^{*'} F + E$$
$$Y = \Psi^{*\prime} G + \Sigma$$

Common factors, loadings and error terms are generated such that final panels X and Y have the following combinations of the dimensions:$T = 50, 100$ and 200 and $N = 20, 50$ and 100. Panels X^* and Y^* are obtained form panels X and Y; to simulate the mixed-frequency pattern of missing observations we omit every second observation in X and Y to achieve 50% omitted observations in panel X^* and Y^*. Similarly, we omit two out of every three observations to achieve 66% distortion of the panel; we omit every three out of four observations to achieve 75% distortion; finally every four out of five observations are missing to achieve 80% distortion in the panel. Next we transfer panels X, X^*, Y, Y^* to stationary form and extract common components from panels X^* and Y^*. The estimated factor $\hat{F}$ is $\sqrt{T}$ times the eigenvectors corresponding to the largest eigenvalue of $XX'$and, therefore, given $\hat{F}$, we can also work out λ ($\lambda = X'F/T$); the same applies to factor G. We use an equation 19 to start a process of non-random initialisation and use an expected maximisation algorithm to re-create panel X. The re-created panel X is compared with the simulation and the conclusion is made regarding how close new panel matches original simulated panel X.

The reported results are based on 1000 repetitions. To evaluate the results we use the *Theils inequality* coefficient popularly referred to as the *Theil's U*. *Theil's U* is a normalised value of a popular loss function, Root Mean Squared Error or RMSE. *Theil's U* has an additional attractive property of having well defined upper and lower bounds; it varies between 0 and 100% where higher results indicate better performance.

where X_{it}^s is simulated panel X and X_{it}^f is panel X re-created using NRI and EM algorithm. We rely on the results of Monte-Carlo simulations to assess the sensitivity of the now-casting performance for Factor-EM Algorithm (henceforth FEMA). We begin the evaluation by assessing the sensitivity of FEMA to the degree of panel distortions. Simulations include only mixed-frequency irregularities because they lead to higher distortions of the panel rather than individually omitted observations. We use this methodology to test panels with T dimensions $\{50, 100, 200\}$ and N dimensions $\{20, 50, 100\}$; we use all combinations of the dimensions. Above we briefly mention the process of inducing missing observations in the panels, however we would like to describe the procedure in detail with the example of the panel X with 50 observations ($T = 50$) and 20 variables ($N = 20$).

The panel X was estimated using equation 16; X is non-stationary and a balanced panel. We omit every second observation in the panel, transforming it to panel X^* size 25×20. Due to this operation half of the simulated panel X observations are omitted, leading to 50% of the panel distortion. This new panel is panel X^* size 25×20 and we transfer it to the stationary format using the formula $\ln(F_1/F_0)$. Panel X^* would be transferred using selection matrix to the panel size 50×20 (panel $X^{(0)}$ in methodology), where every second observation of $X^{(0)}$ equals to zero, reflecting that these observations are unknown. Similarly, we omit every two out of three observations in panel X, such that only every third observation remains unaffected; panel X^* will have a dimension 17×20. Panel $X^{(0)}$ will transfer panel X^* back to the original 50×20 size, but only every third observation in the panel will have a value and the remaining observations are zeros. An example of such transformation will be quarterly GDP data (observed only at the end of every quarter), that is transferred to the monthly frequency. Additionally, we examine the cases when three out of four observations are omitted from panel X; and four out of five observations are omitted resulting in the panel X^* size 10×20. We use equivalent procedure on the panels with different sizes, in particular we look at the combination of the T dimension equals to $\{50, 100, 200\}$ and N dimensions $\{20, 50, 100\}$.

Appendix B, Table XV summarises simulation results by reporting descriptive statistics for the Theil's U (mean and 5, 25, 50, 75 and 95 percentiles). We report results for Factor Expected Maximization and four comparable techniques: Spline, Kalman Filter Interpolation, Factor-Spline, Factor-Kalman Filter. Angilini et al (2006) identify Cubic Spline and Kalman Filters to be reasonable benchmarks for interpolation procedures. We include two types of the techniques: first, we exercise a classical application of both techniques, i.e. direct application of Spline and Kalman Filter to column vectors for interpolation; second, we apply both techniques to the common factor after which we perform the transformation to the panel X using equation 16. The second technique is identified by the expression "Factor- Spline" or "Factor- Kalman Filter".

[Insert table XV around here]

Table XV demonstrates that the variability of the results of Theil's U is the lowest for FEMA in comparison to other approaches; in other words FEMA produces the most consistent results across the simulation. High stability of the parameters is attributable to the uniqueness of the FEMA approach. The methodology of NRI allows precise estimation of preliminary $\hat{F}_t$ values which is key to approximating true F_t. The methodology aims to re-build omitted observations of preliminary $\hat{F}_t$ using available observations

of other highly synchronised common factors $\hat{G}_t$ that we can identify by the high degree of correlation. In methodological section we discuss the methods of identification of panel Y and therefore assume that for the majority of the panel identification of panel Y will be relatively straightforward. We argue that highly synchronised common factors $\hat{F}_t$ and $\hat{G}_t$ should reflect the same common trend. Given that we are able to observe the true path of this common trend (through common factor $\hat{G}_t$) and the degree of distortion of panel $X^*$and therefore the resulting common component $\hat{F}_t^*$ is irrelevant as it synchronised with the observed trend. As an example, we present an extreme case where the common component $\hat{G}_t$ has a perfect correlation ($\rho = 1$) with $\hat{F}_t$. In this case, $\hat{G}_t$ is equivalent to a true value of the higher frequency F_t; therefore, by substituting $\hat{G}_t$ in equation 16 we are able to form panel X_{it} with equivalent frequency to that of panel Y_{it}. It is rare to find perfect correlation between two components, however, examples such as interest rate structures or futures term structure could possibly comply with this condition.

The lowest standard deviation of FEMA holds for all percentiles of results for the distribution of *Theil's U*. This implies that competing approaches are significantly more affected by the change in proportion of omitted observation than the FEMA approach. To illustrate our findings we provide a graphical comparison between two column vectors, one from the true and the other from the FEMA enhanced panel X. The graphs display only the first column vector (true and estimated) since a larger number of columns would be difficult to portray graphically. In doing so, we assume that the first column vector is as good a representation of the interpolation as any other column vector in the panel. This is because it can sensibly demonstrate the relationship between true and estimated vectors. Figures 1, 2, 3 illustrate the FEMA interpolation results; we see that the difference in the graphs is almost indistinguishable, which reaffirms the fact that there is a very small degree of FEMA sensitivity to the omitted observations.

The spline approach produces the largest variation of *Theil's U* results. The classical spline approach has a slightly lower variation of the results than the more advanced Factor-Spline. Such high sensitivity to the proportion of missing observations is explained by the spline methodology, which creates a smooth transition between two values and assumes that all omitted observations lie on the line. Spline is therefore an example of the smoothing function since it is not constructed to identify fluctuations. It collapses to the column vector mean given a large proportion of omitted observations. In other words, it converges to the mean given a large interval between two observable points. Figures 4 and 5 demonstrate the Factor-Spline interpolation; figures 6 and 7 illustrate a classical application of the spline procedure. We observe that both types of spline interpolation have similar results and sensitivity to omitted observations. For panels with 50 percent distortion, the spline methodology is able to produce good results as it will fill

missing observation with the average of two known observations. However, for larger gaps the average of the available observations converges to the mean of the variable, which we observe on figures five and seven. Moreover, the estimated vector demonstrates a large degree of deviation from true observations at the end as a result of spline reverting to its cubic form due to unavailability of final observation (if the last value is not observed).

The Kalman filter demonstrates low mean and standard deviation of *Theil's U* distribution. The results suggest that the Kalman filter is consistently less sensitive, but also has the poorest approximation of Panel X amongst the competing techniques. The Factor-Kalman Filter shows marginally better results than the classic Kalman-Filter approach. The results in Table IV demonstrates a structural "low frequency data" type interpolation. We believe that the methodology of Kalman filters is sub-optimal for this data structure. The problem lies with the filter which has to "learn" the parameters on the low frequency observations; these parameters are subsequently applied by the Kalman-Filter to estimate higher-frequency observations. The parameters vary significantly between low and high frequency to be efficient enough. The Kalman filter performs better during interpolation of ragged edge data, as it learns the parameters on the available data and then it recreates a few missing observations at the end of the sample with the same frequency. Figures 8 and 9 illustrate Factor-Kalman filter; figures 14 and 15 illustrate results of the classical Kalman filter approach. The graphs demonstrate a relatively high volatility of the vector estimations, which, however, never breach the confidence intervals.

Finally, we perform a comparison of FEMA for two types of omitted observations: "ragged edge" against "low frequency". To estimate the accuracy of the FEMA methodology for "ragged edge" panels we perform interpolation for the panel that includes 20% of omitted observations at the bottom of panel X. The methodology is equivalent to "low frequency", such that we estimate preliminary values for common factor $\hat{F}_t$ using high synchronicity with $\hat{G}_t$, then we optimize the results using the EM procedure. Table XV demonstrates that means of *Theil's U* distributions are similar for "low frequency" and "ragged edge" FEMA interpolation procedures. It suggests equivalent effectiveness of interpolation for both structures of omitted variables; however the "ragged -edge" distribution has higher standard deviation, and moreover it is negatively skewed. The results are the reflection of the effect of a longer interval of omitted observations than in "low frequency" structures. It is more challenging to interpolate longer intervals of omitted observations as interpolated values are more likely to drift away from the true path in longer intervals of missing observations; this may lead to error accumulation.

Table XVI demonstrates the sensitivity of the methodology to the sample size with regard to T and N

dimensions. The results show that the sample size of the dataset has only marginal impact on estimations. This result is consistent with Bovin and Ng's (2006) findings that similarly showed marginal changes.

[Insert Table XVI around here]

Table VI FEMA section demonstrates that most significant changes are taking place when we increase sample size along N dimensions. We see that even though the changes are marginal, *Theil's U* demonstrates better approximation to the true panel. Enlarging the sample size along the T dimension produces mixed results, therefore we can argue that better results are achieved for the panels with larger N.

Both spline and Factor-spline interpolation demonstrate low standard deviation of the results while increasing sample size along T and N dimension. Therefore, spline interpolation remains relatively unaffected by changes in samples size. Kalman Filter is the most sensitive among interpolation techniques to the sample size variations along both T and N. We can see a significant increase in the effectiveness of approximations between true and estimated models with an increase in the sample size.

We conclude that FEMA provides better and more consistent interpolation results than the benchmark techniques for the panels with a high degree of distortion. The results are better for the panels with short intervals of omitted observations.The consequences are that ragged-edge panels with long unobserved intervals have a worse interpolation performance than the ones with short intervals. In many cases, ragged-edge panels contain only a few recent unobservable points and, therefore, this issue may be disregarded. The sample size sensitivity analysis confirms consistency of the results.

4.4.2 Empirical application

The empirical application may be viewed as a practical extension of the simulation. The object of the empirical exercise is to demonstrate FEMA methodology application to the real data. To assess the effectiveness of the application we have to compare balanced panel valuies are with interpolated values from artificially distorted panel; this approach will help to assess how well FEMA can interpolate values for real datasets and demonstrate the applicability of the approach. We cannot chose unbalanced data for the empirical application as we won't be able to approximate how close interpolated values to the real ones. Also the choice of unbalanced panels is impossible as application of the factor model approach

is only available to the balanced panels. It would have been interesting to compare factors estimated from balanced and unbalanced panels and make a conclusion regarding the importance of filling missing observations. However, the factor models approach is not applicable to unbalanced panels, therefore we always have to fill missing observations first and later use the principle components estimator. Thus, we believe that an optimal demonstration of the empirical application uses balanced panels, where we introduce a different degree of the distortion according to the objectives of the exercise.

To test FEMA on empirical data we choose two sets of data: the first demonstrates application of the FEMA approach to the panel of mixed-frequency data; the second outlines the case of ragged-edge data and also illustrates the challenges of applying the methodology to the smaller dataset. All panels used in empirical application are balanced data; this condition is mandatory to be able to apply Theil's U function. The first set of data oriented on the datasets where first factor explains less than 50% of variation of the panel; we refer to such datasets as week factor structure panels. The second set of data evaluates datasets where the first factor explains more than 50% of the variation in the panel; we refer to this dataset as a strong factor panel. In the next section we give more detailed description of both data sets and later discuss the results of the empirical application.

4.4.3 Dataset

Both sets of data chosen for empirical application are balanced panels ranging from January 1990 to October 2010 for a total of T=250 monthly observations. The first dataset consists of 120 energy variables which are divided between two panels: a) panel X consisting of 20 oil refinery products prices; b) Panel Y consisting of 100 refinery fundamentals. Panels X and Y are balanced panels consisting of monthly observations that are log differenced and de-trended to suit the requirements of the proposed FEMA algorithm. In particular panel X of from the historical prices of the variety of refinery products, such as gasoline, kerosene, jet fuel, fuel oil and diesel fuel.

The choice of variables for panel Y demonstrates the approach we suggested for the selection of the panel Y in a methodological section. We recommend consideration of the full dataset (120 variables) that was selected using the theoretical justification. Next we can separate datasets on the variables with missing observations (panel X) and variables without missing observations (panel Y). Let's say that the original dataset (120 variables) is intended to be used for forecasting crude oil prices using factor vector autoregressive model. Therefore, the full dataset is intended to be used with the principal component

method, after which factors will be extracted and thus can be used to forecast crude oil prices. We set a challenge for the existing dataset by implying that all but refinery prices data have no missing observations. We would like to interpolate missing observations in refinery prices, such that thereafter we will be able to apply the principle component approach. Therefore, we can divide the original dataset on the panel X, which consist of refinery prises with missing observations; and panel Y which consist of the remaining dataset without missing observations. Note that the dataset used in this exercise is a part of the existing dataset used for FAVAR model in the paper by Zagaglia(2008) and therefore the choice of the variables are justified in the paper. A complete list of the variables can be found in Appendces C and D.

Panel Y shows variables such as the price of crude oil from major exporter countries, which has a direct and very strong correlation with the price of oil refined products. We include information on the amount of the supplied refinery products in barrels per day, and similarly we include information on the amount of refinery product. We include information on the drilling activity that is set to increase available reserves of oil and increase availability of raw material for refineries. We add macroeconomic indicators to reflect demand for oil and refined products. It should be noted that we intended to use GDP growth as a proxy for productivity growth and, therefore, oil consumption (demand) pressure by emerging countries. However, due to quarterly frequency of the GDP we use the industrial production index. We pay significant attention to the US as it is a leading consumer of energy products. We include measures of US monetary aggregates and indicators of confidence. Zhang et al (2008) suggest including US dollar exchange rates since the stability of the US dollar is one of the key elements in understanding oil prices. To provide a proxy for broader financial market sentiments prevailing at the time of the forecasts we include leading stock and bond indices. The dataset reflects the fluctuations of leading oil stocks and OTC derivative spreads. Variables in panel Y are correlated to with refinery prices and should be able to demonstrate major trends in crude oil markets. These trends should be similar to the ones in refinery markets, which will help to interpolate refinery prices.

The second dataset consists of the data from Light Sweet Crude oil spot and future prices for maturities from 1 to 12, for the period January 1990 to October 2010. The first factor extracted from this data explains more than 50% of the variation in the dataset, as the term structure fluctuations demonstrate high correlation. We define this as a dtrong factor structure dataset. We follow three objectives for the application of this dataset: first we investigate the sensitivity of the methodology to a decrease in the panel size. We follow the paper by Bovin and Ng(2006), who show that factor models are only marginally

sensitive to the size of the sample. The second we compare the results of FEMA interpolation for weak and strong factor structures. Third we examine the application of the approach to the ragged edge data, in contrast to the previous case which is applied to the mixed-frequency data. The ragged edge data is that which omits only a few observations at the end or beginning of the dataset. The second dataset is divided on panel X which includes future oil prices from 6 to the 12 month futures; panel Y includes spot prices and 1 to 5 futures. In panel X we artificially induce missing observations, such that the first 70 data points are missing; the rest remains available.

4.4.4 Empirical results

The first step in the FEMA procedure deals with the application of the PC method to the panels X_{it} and Y_{it}. We consider an application to the fully balanced panel with 250 monthly observations for weak factor structures, and then strong factor structure panels with 100 observations. As a result, we extract true $\hat{G}_t$ and $\hat{F}_t$. We consider interpolation for four unbalanced panels with 50, 66, 75 and 80% of irregular data. We also force data irregularities to the balanced panel X and extract unbalanced common factor $\hat{F}_t^*$.

In the empirical literature, there is considerable uncertainty about the appropriate choice of the number of factors since the information criteria seems to provide misleading results in certain cases. For example, Bernanke and Boivin (2003) use three factors for their real-time applications for the US, whereas the Federal Reserve Bank of Chicago publishes a US composite index based on a similar structure but where only one factor of monthly data is chosen. In the application, the number of factors were set equal to one and the rest of the procedures were carried out following the methodology outlined in section 4.3.

[Insert Table XVII around here]

We present the results of the empirical procedure in Table XVII. The interpolation performance of the FEMA is compared to the competing spline and Kalman filter models. We assess the results using mean as a proxy of goodness of fit, and standard deviation to establish efficiency of the methodology. We start by analysing the results for the weak factor model. FEMA demonstrates a higher mean and lower standard deviation of *Theil's U* distribution when varying the proportion of omitted observations in the column vector. The results indicate consistently good approximation with low sensitivity to changes in the proportion of omitted observations.

The spline produces the highest mean for 50% distortion amongst competing approaches. However, the accuracy of approximation severely declines when the proportion of distortion rises. The Kalman Filter demonstrates the extremely poor now-casting ability based on the mean of *Theil's U* distribution. We can see that the quality of now-casting gradually declines as the proportion of missing observations increase. The results are consistent with the results of the simulation exercises.

A strong factor structure improves performance of the FEMA procedure. The performance of splines and KF techniques gradually decline with the increase in the proportion of omitted observations. Poor results of the spline and the Kalman Filter are attributed to the higher volatility of common factors in the strong factor panels, as well as higher volatility of individual vectors. FEMA is able to overcome this problem as factor F_t can model volatility through the volatility of the highly synchronised factor G_t. The results of strong factor models are consistent with those from our simulation exercise. Overall, panels with a strong factor structure have improved results over those from panels with weaker factor structures when using the FEMA methodology. However, this improvement is generally marginal.

We perform an additional evaluation of the empirical validity of the FEMA approach using the assessment of the structural parameters of the panel regression. We regress a panel Y which consists of eight variables including refinery prices of motor gasoline, aviation gasoline, kerosene-type jet fuel, kerosene, fuel oil, diesel fuel, propane, residual fuel oil. The regressors are summarised in panel X which consists of the costs of oil imported from Mexico, Nigeria, Venezuela, and average oil prices from OPEC and Non- OPEC countries. Original panels X and Y contained 100% of observations with monthly frequencies; to perform the exercise we decreased both panels samples on 50%, 66%, 75% and 80% by methodological deletion of the observations. For example for the panel with sample equals to 50% from the original, we delete every second observation; for the sample with 66% missing observations we keep every third observation and delete two out of every three; for panels with 75% missing observations we keep every forth observation and for the panels with 80% missing observations we keep every fifth observation in the panel. We regress low frequency panel Y on low frequency panel X; the results are given in the Table XVIII.

In the second step we use FEMA methodology to interpolate low frequency panels Y^* and X^* with 50%, 66%, 75% and 80% back to the 100% monthly observations. We regress interpolated panel Y on panel X and display the coeffcients in the Table XVIII. In the table we present coeffcients and standard errors, which help to assess structural parameters. We can see that the parameters of the high frequency regressions are relatively stable and the values are close between the parameters from the panels inter-

polated from 50%, 66%, 75% and 80% panels of missing observations. At the same time the parameter for the low frequency panel is significantly different between them and substantially different from high frequency parameters. The difference between structural parameters is due to the variations in the sample size; while size of T dimension in the 100% balanced panel is 240 observations; the sample size in the panel with 80% missing observations is 48. Such a drastic difference in the sample sizes of the regression leads to the difference in the structural parameters. At the same time we can see that parameters which are interpolated to have the same sample size (the regressions between panels Y and X) have very similar structural parameters, which confirms the theory. To conclude, we can observe that FEMA interpolation methodology panels generate stable and consistent observations irrespective of the proportion of missing observations. This is confirmed by the structural parameters, which are similar for the panels interpolated from panels with 50% to 80% missing observations.

4.5 Conclusions

Our study is inspired by the problem of irregularity of data observations, which is more than usual in the large dimensional datasets. We aimed to create a comprehensive method that can efficiently recreate numerous patterns of omitted observations between cross-sectionally correlated time series variables. We accomplished the goal by designing into Factor Expected Maximization Approach that can successfully provide substitutes for omitted observations in the column vectors while taking into consideration cross-sectional correlation within the panel.

The methodology of FEMA builds on factor model theory and more precisely on diffusion index methodology that suggests using common trends between variables to provide omitted observations. The methodology is linked directly to the principal components method that helps to extract common trends from lower and higher frequency data. The higher frequency trends are used as a stencil to substitute preliminary estimations to the omitted observations; expected maximisation technique optimises the results. The methodology differs from random initialisation that is common among EM applications. Non-random initialisation allows estimation of more precise preliminary values than random initialisation and therefore, mitigates sensitivity to the long intervals and/or large proportions of omitted variables. The conclusion is confirmed by the results of simulation and empirical exercises. According to simulation exercises, the competing models failed to provide consistent outperformance when we induced large intervals of omitted variables. In addition, we ensured that the results hold for both large and medium

size panels by performing size sensitivity evaluation. The empirical application used energy data to demonstrate the methodology "in action". We used both strong and weak factor model structures; and as expected, strong factor FEMA had better results.

Let us mention that FEMA has limitations and opportunities for further improvements. First, non-random initialisation is only available for the factor base panel, which can be matched with the identical high frequency common trend. We have the potential to mitigate this issue by splitting original large dimensional panel into two parts: omitted observations X and balances panel Y. A factor structure of large dimensional panels implies that they share a common trend along the panel. By splitting the original panel into two we preserve factor structure, and allow the balanced panel to act an approximator of the second part. This scenario is a realisticoccurrence in financial markets (for example, heating oil term structure).

Second, the methodology allows only for variations in the common trend, assuming that factor loading remains stable. Although we believe that this assumption is trustworthy, we feel that the research can be extended to include flexible factor loading that varies from low to high frequency. Flexible factor loading can be estimated using a similar methodology for non-random initialisation and EM. Third, we believe that the examination of the methodology is limited solely to monthly-frequency based applications of simulations and empirical study. The application can be extended to weekly, daily and hourly frequencies, though higher rates are subject to further investigation.

The results of the simulations and empirical exercises suggest that it is a difficult task to provide omitted observations on such a large scale as 50 percent (and more) omitted observations in large dimensional data. However, the results demonstrate that it is possible to identify omitted observations with around 60 percent precision, which is shown by Theil's U results. The results are comparable with academic literature on now-casting and large dimensional interpolation such as Stock and Watson (2002a), Schumacher and Breitung (2008). Additionally, we leave the room for further investigation and improvement of the method, as well as wider empirical applications.

References

[1] Appendix A: Proofs

Lemma 2 *Let assumptions A-H from Bai(2004), then:*

(i) $\frac{1}{T^2}\sum_{t=1}^{T} G_t F_t = \rho + O_p(\frac{1}{T})$

(ii) $\frac{1}{T^2}\sum_{t=1}^{T} G_t(\hat{F}_t - F_t) = \frac{1}{T}(O_p(\frac{1}{T}) + O_p(\frac{1}{\sqrt{N}}))$

(iii) $\frac{1}{T^2}\sum_{t=1}^{T} F_t(\hat{G}_t - G_t) = \frac{1}{T}(O_p(\frac{1}{T}) + O_p(\frac{1}{\sqrt{N}}))$

(iv) $\frac{1}{T^2}\sum_{t=1}^{T}(\hat{G}_t - G_t)(\hat{F}_t - F_t) = O_p\left(\frac{1}{T\delta_{NT}^2}\right),$

where $\delta_{NT}^2 = \min\left\{N, T^2\right\}$

Proof. Proof. Lemma 2 ■

We begin with term *(i)* which we re-write accordingly:

$$\frac{1}{T^2}\sum_{t=1}^{T} G_t F_t = \rho\frac{1}{T^2} G_t G_t' + \frac{1}{T^2}\sum_{t=1}^{T} G_t e_t$$

By Central Limit Theorem $\hat{G}_t\hat{G}_t/T^2 = I$ and by Functional Central Limit Theorem $\frac{1}{T^2}\sum_{t=1}^{T} G_t F_t = \rho + O_p(T)$. Following *Bai(2004) Lemma B.4(i)* we determine both terms *(ii)* and *(iii)* to be $T^{-1}(O_p(T^{-1}) + O_p(N^{-1/2}))$. Finally, we are using Cochy-Schwarts inequality and later applying *Bai(2004) Lemma B.1* to express term *(iv)* as below:

$$\frac{1}{T^2}\sum_{t=1}^{T}(\hat{G}_t - G_t)(\hat{F}_t - F_t) \leq \left[\frac{1}{T^2}\sum_{t=1}^{T}(\hat{G}_t - G_t)^2\right]^{1/2}\left[\frac{1}{T^2}\sum_{t=1}^{T}(\hat{F}_t - F_t)^2\right]^{1/2} = O_p\left(\frac{1}{T\delta_{NT}^2}\right)$$

this is due to $\frac{1}{T}\sum_{t=1}^{T}\left\|\hat{F}_t - F\right\|^2 = O_p(\frac{1}{\delta_{NT}^2})$ and δ_{NT}^2 defined in *Bai(2004)* $\delta_{NT}^2 = \min\left\{N, T^2\right\}$.

Proof. Proposition 2 ■ ■

Follows from equation 19 $\hat{\rho} = \left[\sum_{t=1}^{T} \hat{G}_t \hat{G}_t\right]^{-1} \left[\sum_{t=1}^{T} \hat{G}_t \hat{F}_t\right]$. We substitute $(\hat{G}_t - G_t)$ and $(\hat{F}_t - F_t)$ to arrive at equation 22 :

$$\begin{aligned} \hat{\rho} &= \left[\frac{1}{T^2}\sum_{t=1}^{T} \hat{G}_t \hat{G}_t\right]^{-1} \left[\frac{1}{T^2}\sum_{t=1}^{T} (G_t + (\hat{G}_t - G_t))(F_t + (\hat{F}_t - F_t))\right] \\ &= \frac{1}{T^2}\sum_{t=1}^{T} G_t F_t + \frac{1}{T^2}\sum_{t=1}^{T} G_t(\hat{F}_t - F_t) + \frac{1}{T^2}\sum_{t=1}^{T} F_t(\hat{G}_t - G_t) + \frac{1}{T^2}\sum_{t=1}^{T} (\hat{G}_t - G_t)(\hat{F}_t - F_t) \end{aligned} \tag{22}$$

Equation 22 is true while $\sum_t \hat{G}_t \hat{G}_t / T^2 = I_k$ holds. Next using results from *Lemma* 2 we establish:

$$\hat{\rho} = \rho + O_p(\frac{1}{T}) + O_p(\frac{1}{T^2}) + O_p(\frac{1}{T\sqrt{N}}) + O_p(\frac{1}{T^2}) + O_p(\frac{1}{T\sqrt{N}}) + O_p(\frac{1}{T\delta_{NT}^2})$$

$$T(\hat{\rho} - \rho) = O_p(1) + O_p(\frac{1}{\sqrt{N}}) \tag{23}$$

where $\delta_{NT}^2 = \min\left\{N, T^2\right\}$. In 23 we left with two slowest converging terms.

Proof. Theorem 2

Consider equation 19, where $\hat{\rho}$ converges to the true value with the rate established in 23.

$$T(\hat{\rho} - \rho) = O_p(1) + O_p(1/\sqrt{N}) = I + II$$

Consider I in light of Proposition 2 and Lemma 2:

$$\begin{aligned} I &= T[\sum_{t}^{T} H_G G_t G_t' H_G']^{-1}[\sum_{t=1}^{T} H_G G_t e_t] = \\ &= [\frac{1}{T^2}\sum_{t=1}^{T} H_G G_t G_t' H_G']^{-1}[\frac{1}{T}\sum_{t=1}^{T} H_G G_t e_t] = a \times b = O_p(1) \end{aligned}$$

According to *Assumption A(2)in Bai(2004) p.140* (see section 2.3.1)and *Assumption H in Bai(2004)*

*p148 (*see section 6*):*

$$a = \frac{1}{T^2}\sum_{t=1}^{T} H_G G_t G_t' H_G' \xrightarrow{d} H_G \int_0^1 B_u B_u' H_G'$$
$$b = \frac{1}{T}\sum_{t=1}^{T} H_G G_t e_t \xrightarrow{d} H_G \int_0^1 B_u dB_e$$

B_u andB_e are vectors of Brownian motions, where B_u defined in *Bai(2004), p140* and presented in section 2.3.1; B_e is a motion of e_t defined in B*ai(2004), p148* and presented in details in section 6. B_e has variance $\sigma_e^2 = \lim_{T\to\infty} Var[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} e_t]$,such that $B_e(r) \overset{d}{=} \sigma_e B(r)_{st} \sim \sigma_e N(0,r)$; and rotation matrix H is defined in 3.4. As a result

$$a \times b \xrightarrow{d} [H_G \int_0^1 B_u B_u' H_G']^{-1}[\sigma_e H_G \int_0^1 B_u dB_{st}]$$

Therefore, for the case when $1/\sqrt{N} \to 0$ limiting distribution for correlation coeffisient is defined:

$$T(\hat{\rho} - \rho) \overset{d}{=} \sigma_e [H_G \int_0^1 B_u B_u' H_G']^{1/2} N(0,1)$$

■

Lemma 3 *: Under assumptions A-G from Bai(2003) and (N,T)→ ∞, then* :

(i) $\frac{1}{T}\sum_{t=1}^{T} G_t F_t = \rho + O_p(\frac{1}{\sqrt{T}})$

(ii) $\frac{1}{T}\sum_{t=1}^{T} G_t(\hat{F}_t - F_t) = O_p(\frac{1}{\gamma_{NT}^2})$

(iii) $\frac{1}{T}\sum_{t=1}^{T} F_t(\hat{G}_t - G_t) = O_p(\frac{1}{\gamma_{NT}^2})$

(iv) $\frac{1}{T}\sum_{t=1}^{T} (\hat{G}_t - G_t)(\hat{F}_t - F_t) = O_p(\frac{1}{\gamma_{NT}^2})$

where, $\gamma_{NT}^2 = \min\{N,T\}$

Proof. Lemma 3

Consider equation 19, we re-write all terms starting from *(i):*

$$\frac{1}{T}\sum_{t=1}^{T} G_t F_t = \rho \frac{1}{T} G_t G_t' + \frac{1}{T}\sum_{t=1}^{T} G_t e_t.$$

We use Law of Large numbers for stationary process to determin that $\hat{G}_t\hat{G}_t/T = I$, next we use Central Limit Theorem find rate of convergence for term *(i)*that is: $\rho + O_p(\frac{1}{\sqrt{T}})$.By direct applicaton of *Bai(2003) Lemma B.2 p.164-165* we determine terms *(ii)* and *(iii)* to be $O_p(\frac{1}{\gamma_{NT}^2})$, where $\gamma_{NT}^2 = \min\{N, T\}$. Finally, we use Cauchy-Schwartz inequality and *Bai(2003), Lemma A.1 p. 158* to be able to express term (iv) as follow:

$$\frac{1}{T}\sum_{t=1}^{T}(\Delta\hat{G}_t - \Delta G_t)(\Delta\hat{F}_t - \Delta F_t) \leq \left[\frac{1}{T}\sum_{t=1}^{T}(\Delta\hat{G}_t - \Delta G_t)^2\right]^{1/2}\left[\frac{1}{T}\sum_{t=1}^{T}(\Delta\hat{F}_t - \Delta F_t)^2\right]^{1/2} = O_p(\frac{1}{\gamma_{NT}^2})$$

■

Proof. Proposition 3 According to equation 19 then $\hat{\rho}$ can be expressed as follow:

$$\hat{\rho} = \left[\sum_{t=1}^{T}\hat{G}_t\hat{G}_t\right]^{-1}\left[\sum_{t=1}^{T}\hat{G}_t\hat{F}_t\right]$$

In proof of lemma 3 we establish that $\sum_t \hat{G}_t\hat{G}_t/T = I_k$. Next we substitute $(\hat{G}_t - G_t)$ and $(\hat{F}_t - F_t)$ and as a result arive to equation 24.

$$\begin{aligned}\hat{\rho} &= \left[\frac{1}{T}\sum_{t=1}^{T}\hat{G}_t\hat{G}_t\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}(G_t + (\hat{G}_t - G_t))(F_t + (\hat{F}_t - F_t))\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}G_tF_t + \frac{1}{T}\sum_{t=1}^{T}G_t(\hat{F}_t - F_t) + \frac{1}{T}\sum_{t=1}^{T}F_t(\hat{G}_t - G_t) + \frac{1}{T}\sum_{t=1}^{T}(\hat{G}_t - G_t)(\hat{F}_t - F_t)\end{aligned} \tag{24}$$

By applying lemma 3, we have:

$$\begin{aligned}\hat{\rho} &= \rho + O_p(\frac{1}{\sqrt{T}}) + O_p(\frac{1}{\gamma_{NT}^2}) \\ \sqrt{T}(\hat{\rho} - \rho) &= O_p(1) + O_p(\frac{\sqrt{T}}{N})\end{aligned} \tag{25}$$

■

Proof. Theorem 3

Based on equation 19 and convergence rate from 25 we establish limiting distribution of $\hat{\rho}$ for stationary set of estimators $\hat{F}$ and $\hat{G}$

$$\sqrt{T}(\hat{\rho}-\rho) = O_p(1) + O_p(\frac{\sqrt{T}}{N}) = I + II$$

Consider I

$$\begin{aligned} I &= \sqrt{T}[\sum_{t=1}^{T} H_G G_t G_t' H_G']^{-1}[\sum_{t=1}^{T} H_G G_t e_t] = \\ &= [\frac{1}{T}\sum_{t=1}^{T} H_G G_t G_t' H_u']^{-1}[\frac{1}{\sqrt{T}} H_G G_t e_t] = a \times b = O_p(1) \end{aligned}$$

According to *Assumption A Bai(2003) p.141 and Assumption F(4) p.144*

$$\begin{aligned} a &= \frac{1}{T}\sum_{t=1}^{T} H_G G_t G_t' H_G' \xrightarrow{p} \Sigma_G \\ b &= \frac{1}{\sqrt{T}} H_G G_t e_t \xrightarrow{d} N(0, \Phi_i) \end{aligned}$$

where Σ_G is positive definite matrix $r \times r$ and $\Phi_i = p\lim_{T\to\infty}(1/T)\sum_{t=1}^{T} E[G_t^0 G_s^{0\prime} e_{is} e_{it}]$. For more details see *Bai(2003).*

$$a \times b \xrightarrow{d} \Sigma_G^{-1}\Phi_i^{1/2} N(0,1)$$

Therefore, for the case when $\sqrt{T}/N \to 0$ limiting distribution of correlation coeffisient for stationary panels is:

$$\sqrt{T}(\hat{\rho}-\rho) = \Sigma_G^{-1}\Phi_i^{1/2} N(0,1)$$

■

Table XV. Simulation results

	Factor Expected Maximization Approach						Ragged Edge					
	Mean	0.05	0.25	0.5	0.75	0.95	Mean	0.05	0.25	0.5	0.75	0.95
50%	58.9712	51.0434	56.4722	58.7229	61.2324	66.2153	56.2508	40.4791	45.7541	60.2757	62.0093	69.7003
66%	58.92656	51.0857	56.5606	58.7165	61.2608	66.2245						
75%	58.8574	51.4755	56.6077	58.6495	61.2661	66.2273						
80%	56.6576	51.2521	56.4434	56.617	61.122	66.1377						
	Factor-Spline						Spline					
50%	50.2519	44.1168	47.871	50.5048	52.451	55.7573	50.377	44.6167	48.2761	50.4681	52.5331	55.5913
66%	45.8117	35.2747	37.6823	49.0206	49.7237	44.0145	47.31358	41.3817	43.4565	50.2728	50.9662	42.8001
75%	36.644	31.3700	33.9908	37.1537	39.2925	43.5253	36.5248	31.3205	34.1659	32.2884	36.9412	42.1295
80%	31.9167	27.0375	30.6405	33.6544	38.1356	40.6587	35.7224	29.5923	38.9675	36.7984	37.5991	41.3295
	Factor-Kalman						Kalman Filter					
50%	34.8816	21.7557	32.4377	36.3804	39.8356	42.5878	33.921	20.0006	28.3463	34.2824	38.2702	54.9931
66%	34.43819	22.8701	32.3789	35.9149	39.3451	40.0747	33.46904	19.9205	27.7763	32.4243	36.7178	50.6164
75%	33.6383	25.8827	31.0504	33.5934	36.9166	38.8057	30.5478	19.5609	25.3903	31.3626	33.7893	45.3792
80%	31.8674	25.6389	30.9377	32.543	35.6305	37.6174	28.5812	18.3684	24.5167	31.1808	33.4696	44.5799

Table XV shows reversed and normalised RMSE loss functions, such that the functions are allowed to take values between 0 and 100%. The results are equal to 100% when all filled observations exactly match true values of these observations. The first column in the table represent average results achieved by competing methodologies; the remaining columns represent quantiles of the distribution; the results are estimated over 1000 simulations. The results are estimated for different number of missing observation in the panel (from 50% to 80%).

Table XVI. Sensitivity Analysis

	50% omitted			66% omitted			75% omitted			80% omitted		
Factor Expected Maximization Approach												
	N=20	N=50	N=100	N=20	N=50	N=100	N=20	N=50	N=100	N=20	N=50	N=100
T=50	58.5179	58.7572	62.3124	58.5350	58.3603	62.4165	57.6310	57.8380	61.4553	56.3574	56.5967	60.1519
T=100	59.2763	61.2484	60.9687	58.7659	60.4370	60.5583	58.9541	60.8952	60.5984	56.6576	58.6297	58.3500
T=200	57.4127	59.7979	60.4372	59.3838	56.7916	56.3663	59.8861	62.2389	62.9313	57.6183	55.2331	54.5938
Factor Spline												
T=50	49.8648	49.5208	49.3760	42.3684	42.0280	41.8805	35.4416	35.1012	34.9537	34.3461	34.0021	33.8573
T=100	50.9819	50.4422	50.4436	42.8170	42.6202	42.2346	36.6404	36.1430	36.1570	36.2253	35.6856	35.6870
T=200	50.4112	50.2422	49.7874	43.8533	44.0430	44.4915	37.4187	37.3280	36.8804	37.4541	37.6231	38.0779
Factor Kalman Filter												
T=50	32.1285	34.4159	35.0171	33.4047	35.7785	36.3905	32.3606	34.9342	35.5462	30.1782	32.4656	33.0668
T=100	34.0280	37.1530	37.7717	34.2311	37.3813	37.9685	33.6833	36.7336	37.3208	31.6524	34.7774	35.3961
T=200	36.3003	38.2401	39.1725	34.5543	32.5569	31.6326	35.6740	37.5724	38.4967	33.6158	31.6760	30.7436
Spline												
T=50	49.1000	49.1618	48.8360	38.4134	38.5310	38.1737	34.6937	34.7114	34.4531	34.3497	34.4115	34.0857
T=100	50.8888	50.6519	50.5253	38.7426	38.5732	38.4070	36.5487	36.2798	36.1136	35.2624	35.0255	34.8989
T=200	50.8100	50.4753	50.2691	40.2751	40.5225	40.7845	38.0254	37.7780	37.5160	36.6545	36.9892	37.1954
Kalman Filter												
T=50	31.5932	33.3847	34.4739	32.5144	34.3293	35.4212	28.9807	30.7956	31.8875	29.6452	31.4367	32.5259
T=100	33.3098	35.2482	36.3478	33.6619	35.6453	36.7413	30.5784	32.4628	33.5588	29.9393	31.8777	32.9773
T=200	34.3492	37.8282	38.2832	33.9325	30.4427	30.0921	31.2258	34.7156	35.2651	31.3794	27.9004	27.4454

Table XVI shows reversed and normalised RMSE loss functions, such that the functions are allowed to take values between 0 and 100%. The results are equal to 100% when all filled observations exactly match true values of these observations. Table shows average results achieved by competing methodologies; the results are estimated over 1000 simulations for the panel with different number of columns N and observations T. The results are estimated for different number of missing observation in the panel (from 50% to 80%).

Table XVII Empirical results

Weak factor structure				
N=20/T=100	50% omitted	66% omitted	75% omitted	80% omitted
Factor	56.0694	56.0467	55.7485	55.4007
Factor-Spline	61.7721	48.0266	49.5426	43.1693
Factor-Kalman	49.292	35.1382	34.4789	35.1047
SplineIndiv	67.5456	48.3019	54.1855	47.1131
KalmanIndiv	42.51	41.4901	40.3793	40.4729

Strong factor structure				
N=20/T=100	50% omitted	66% omitted	75% omitted	80% omitted
Factor	60.0781	59.9577	59.9975	60.6792
Factor-Spline	47.6854	45.3498	38.8739	42.2862
Factor-Kalman	36.5646	35.9756	21.144	36.3239
SplineIndiv	48.1778	46.4598	39.7355	41.9037
KalmanIndiv	32.7832	32.1087	30.9853	31.439

Table XVII shows reversed and normalised RMSE loss functions, such that the functions are allowed to take values between 0 and 100%. The results are equal to 100% when all filled observations exactly match true values of these observations. Table shows average results achieved by competing methodologies while filling missing observations for empirical dataset. The results are estimated for different number of missing observation in the panel (from 50% to 80%). Additionally, we demonstrate a variation of the results for two types of panels: panels with strong factor structure and panels with weak factor structure.

Table XVIII Empirical results

	High frequency				Low frequency			
	50%	66%	75%	80%	50%	66%	75%	80%
Oil import cost, Mexico	0.3703	0.3658	0.2546	0.3368	-0.0850	0.0000	-0.3201	-0.0506
	0.0832	*0.0879*	*0.0886*	*0.1012*	*0.1215*	*0.0002*	*0.1485*	*0.1850*
Oil import cost, Nigeria	0.2568	0.2325	0.3654	0.1599	0.3054	0.2472	0.6579	0.5851
	0.0674	*0.0714*	*0.0719*	*0.0822*	*0.0978*	*0.0990*	*0.1175*	*0.1638*
Oil import cost, Venezuela	0.0730	0.1217	-0.0268	0.1518	0.0931	0.4409	-0.0531	-0.2162
	0.0495	*0.0524*	*0.0528*	*0.0603*	*0.0761*	*0.1166*	*0.1107*	*0.0973*
Oil price OPEC	-0.0019	0.0527	-0.1016	-0.1215	-0.4271	0.0477	-0.2608	-0.0652
	0.0901	*0.0953*	*0.0961*	*0.1098*	*0.1271*	*0.0791*	*0.1731*	*0.1437*
Oil price Non-OPEC	0.8957	0.8656	1.1254	1.0794	0.4964	-0.3333	0.4571	0.3199
	0.1091	*0.1154*	*0.1163*	*0.1329*	*0.1624*	*0.1353*	*0.2043*	*0.2493*
Constant	0.0030	0.0028	0.0014	0.0000	-0.0009	-0.0018	-0.0001	0.0000
	0.0004	*0.0008*	*0.0008*	*0.0009*	*0.0012*	*0.0019*	*0.0015*	*0.0021*
R-sruared	0.7186	0.7057	0.6968	0.6448	0.2342	0.2835	0.6168	0.6156

Table XVIII demonstrates parameters, standard errors and the goodness of fit of the regressions run using variables with smaller number of observations due to the fact that some of the proportion of these variables are missing (low frequency). The proportion of missing observation fluctuates between 50% and 80% if we establish original number of observations equals to 100%. I interpolate these variables using proposed methodology and perform the regression; the resulting parameters are reported in the columns marked “high frequency”.

Appendix C. List of figures

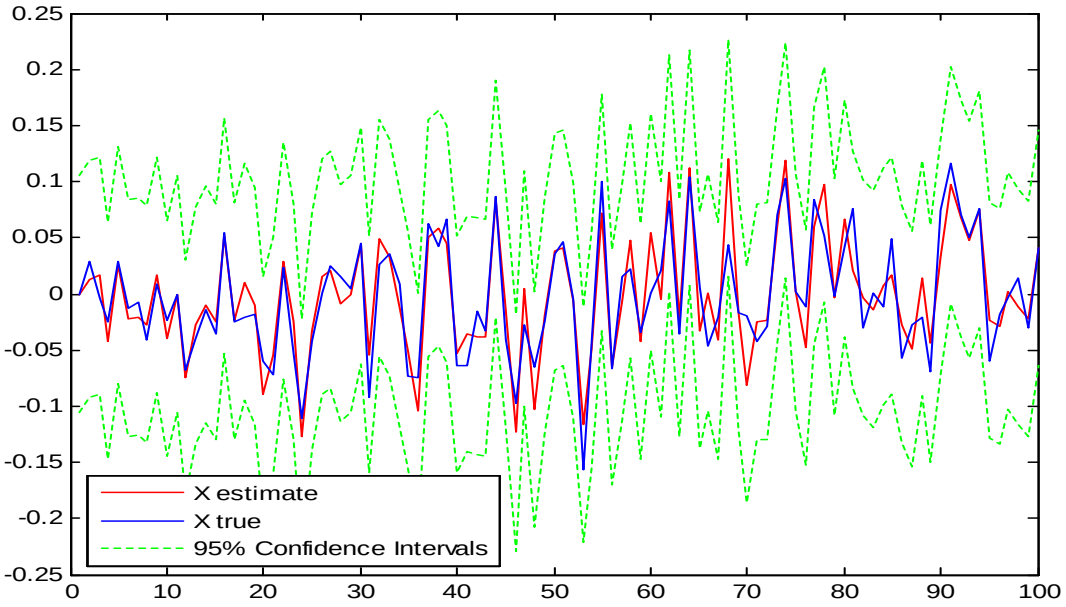

Figure 1: The graph of FEMA interpolation, for dataset with 50 % omitted observations

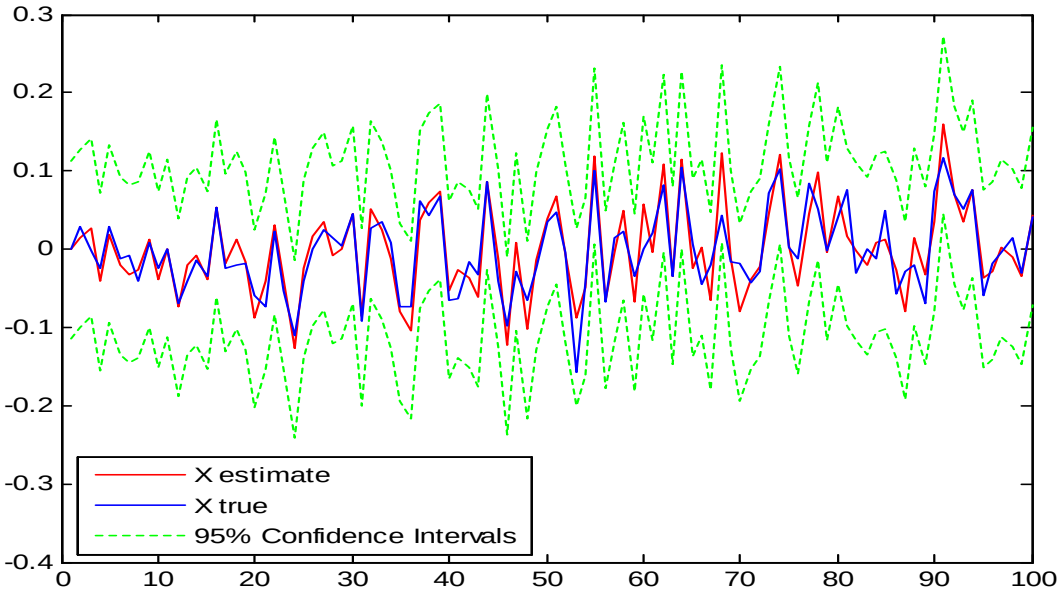

Figure 2: The result of FEMA interpolation, for dataset with 75 % omitted observations

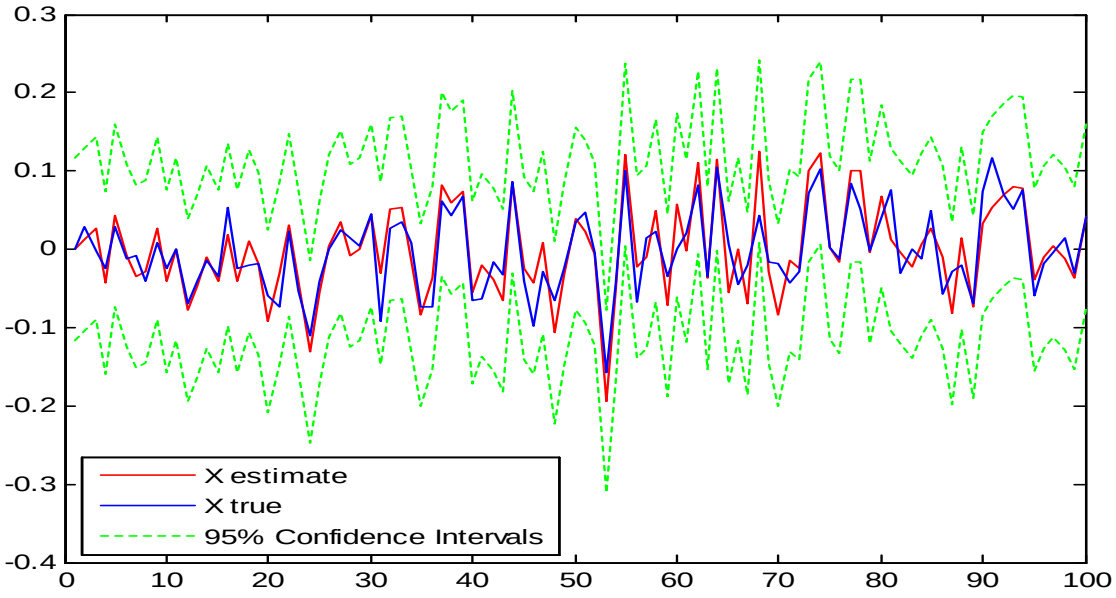

Figure 3: The results of FEMA interpolation, for dataset with 80% omitted omitted observations

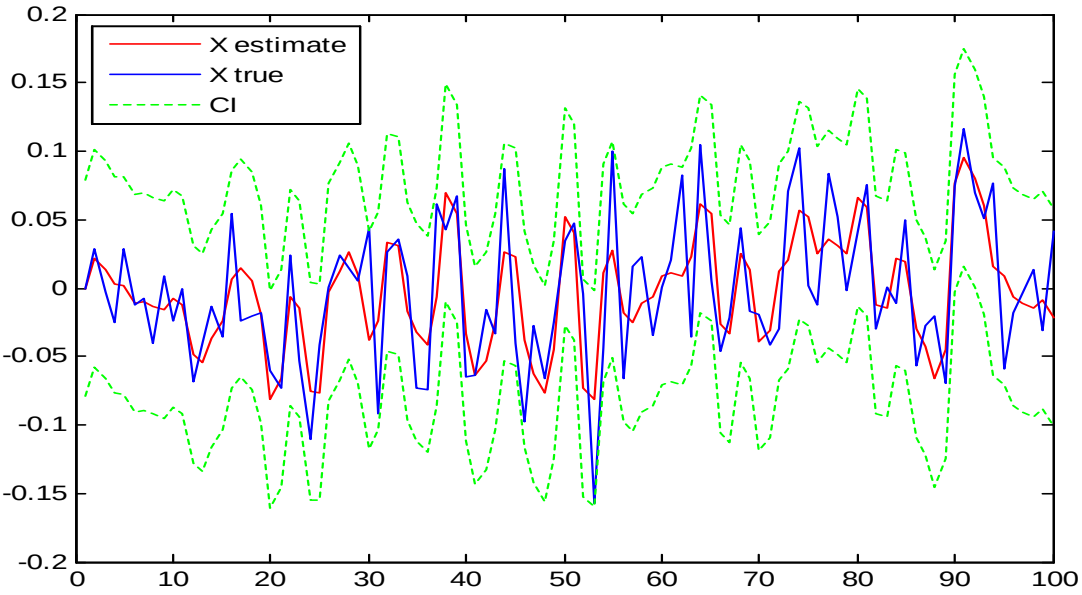

Figure 4: The results of Factor Spline, for datasets with 50% omitted observations

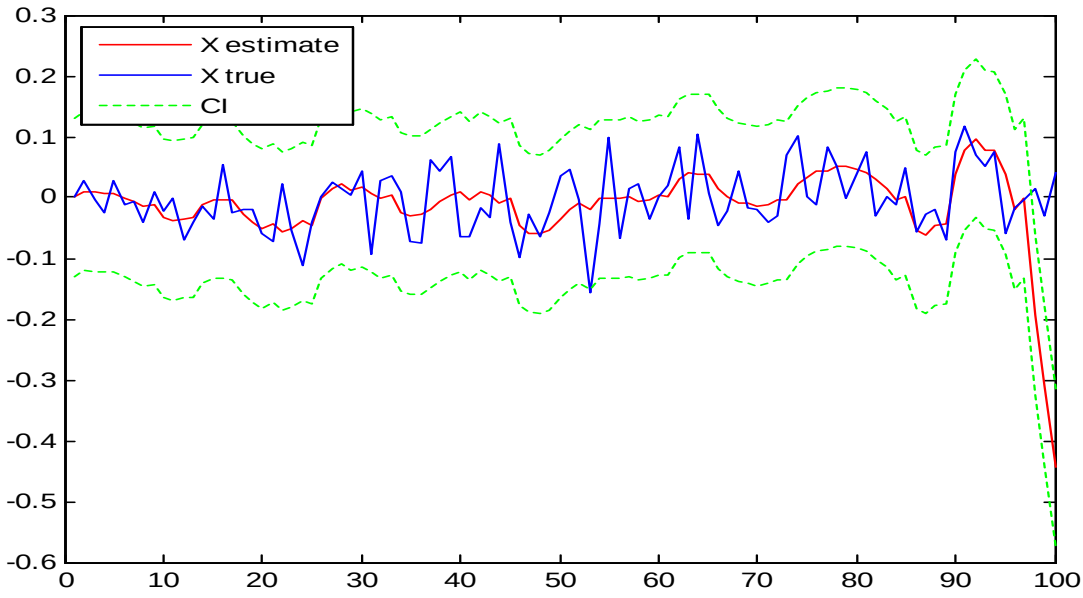

Figure 5: The results of Factor Spline, for datasets with 75 % omitted observations

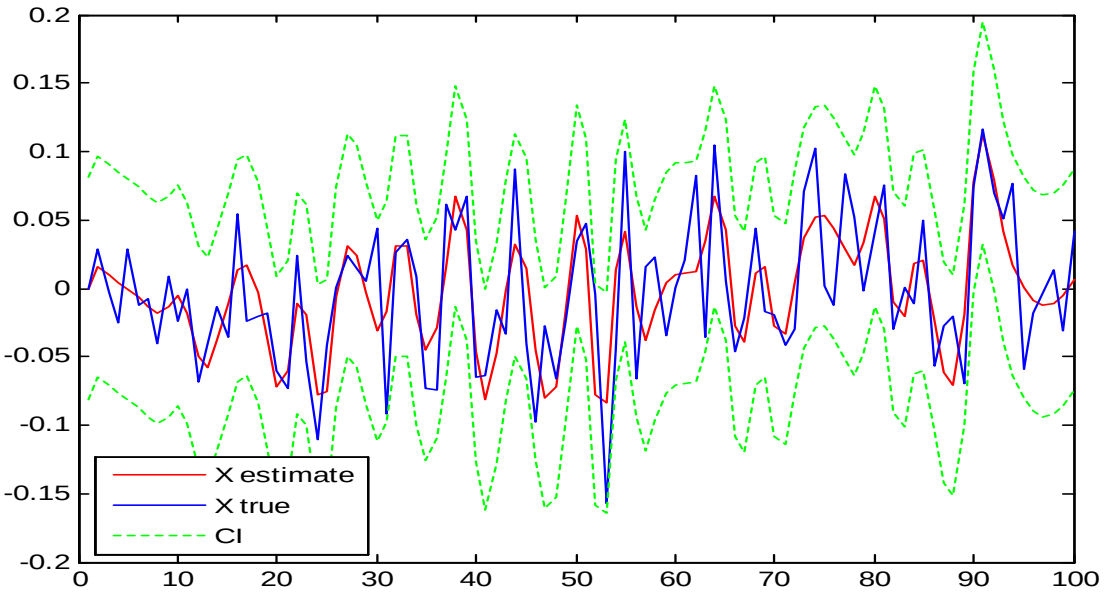

Figure 6: The results of Spline, for datasets with 50 % omitted observations

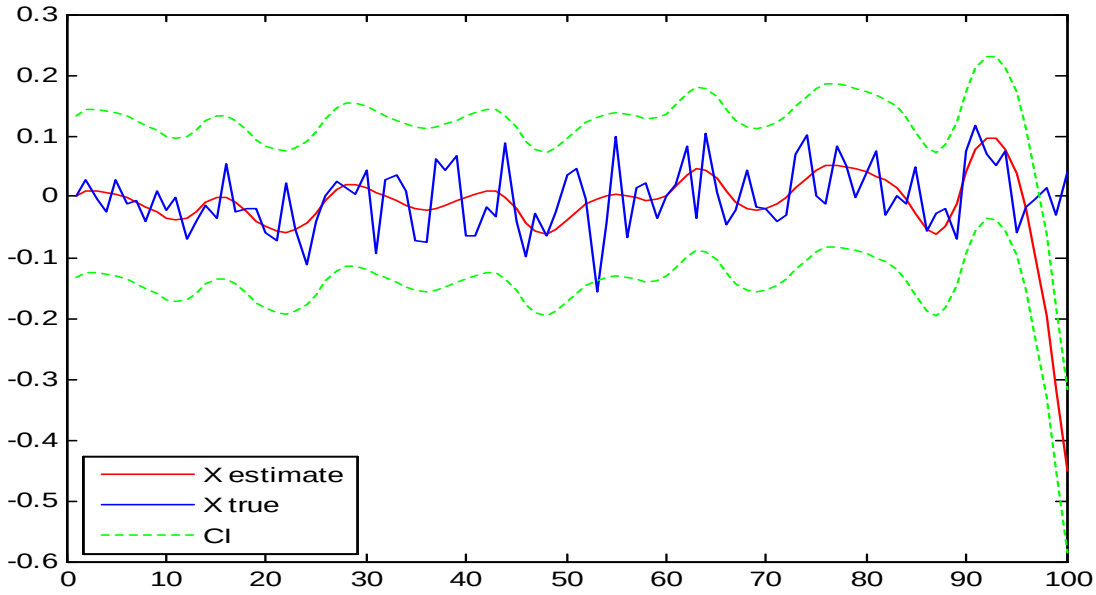

Figure 7: The results of Spline, for datasets with 75 % omitted observations

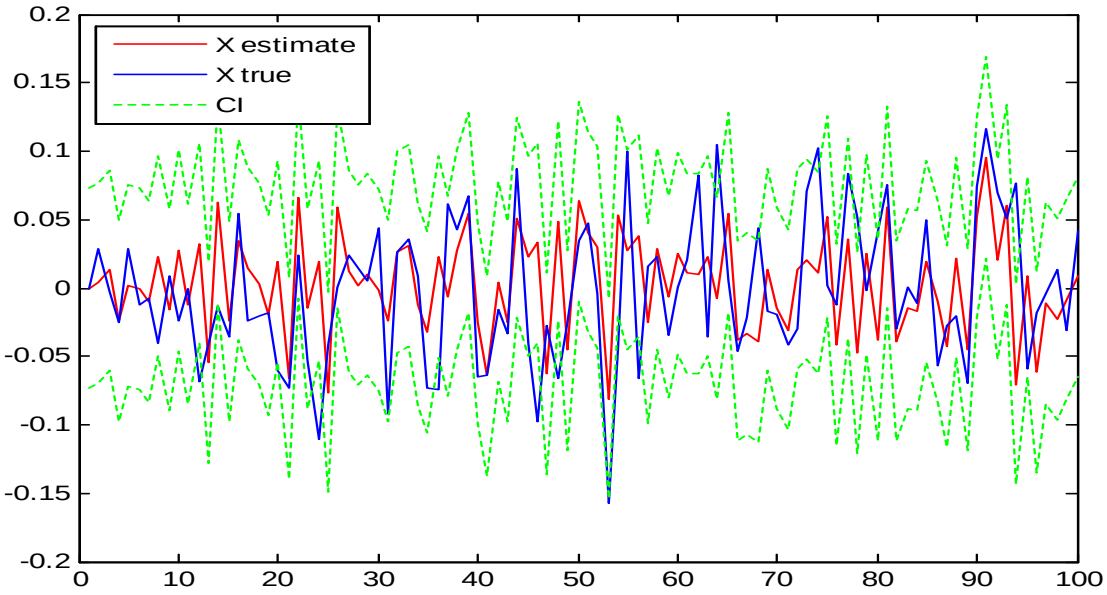

Figure 8: The results of Factor Kalman Filter, for datasets with 50 % omitted observations

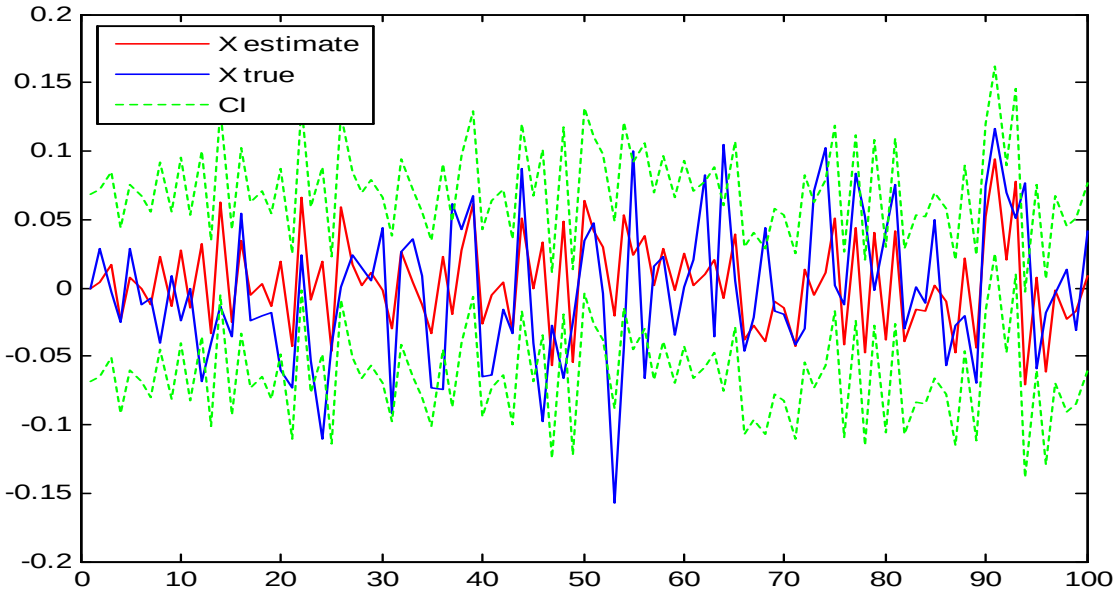

Figure 9: The results of Factor Kalman Filter, for datasets with 75 % omitted observations

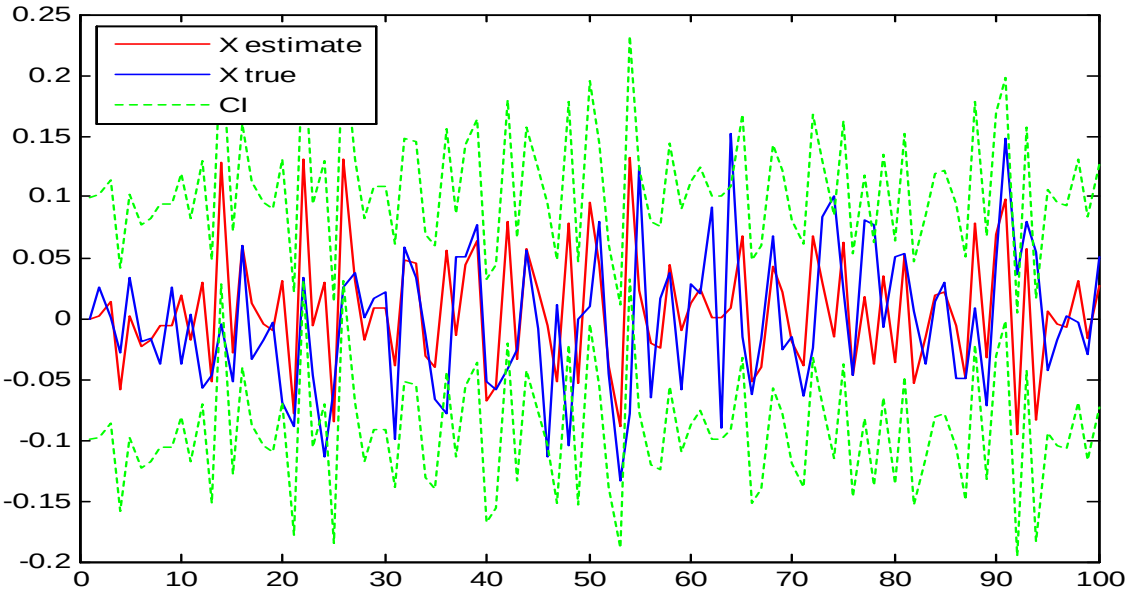

Figure 10: The results for Kalman Filter, for datasets with 50 % omitted observations

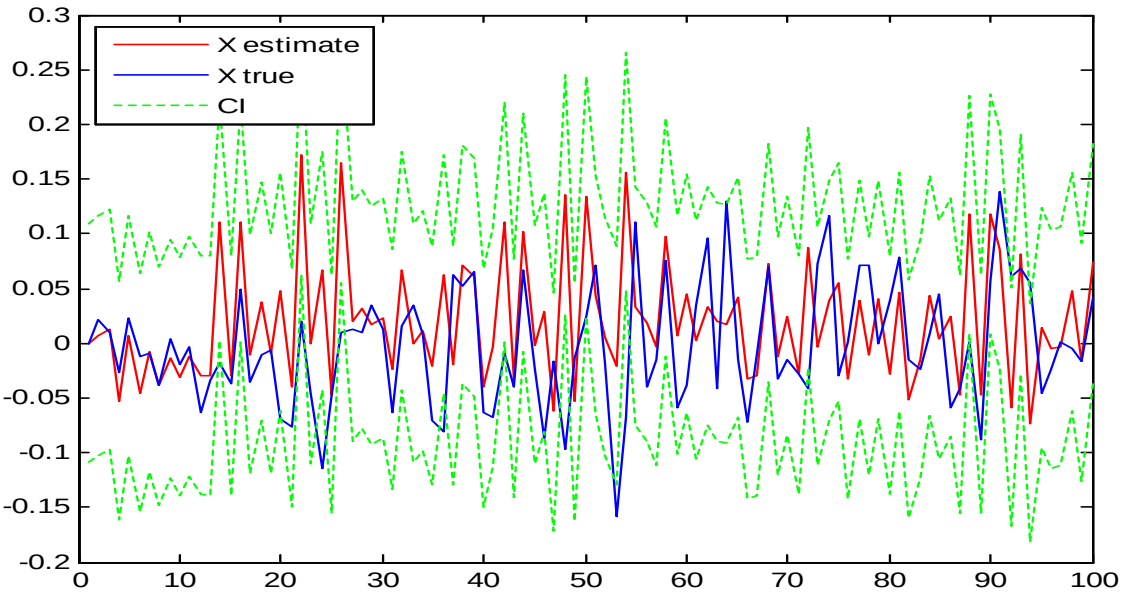

Figure 11: The results for Kalman Filter, for datasets with 75 % omitted observations

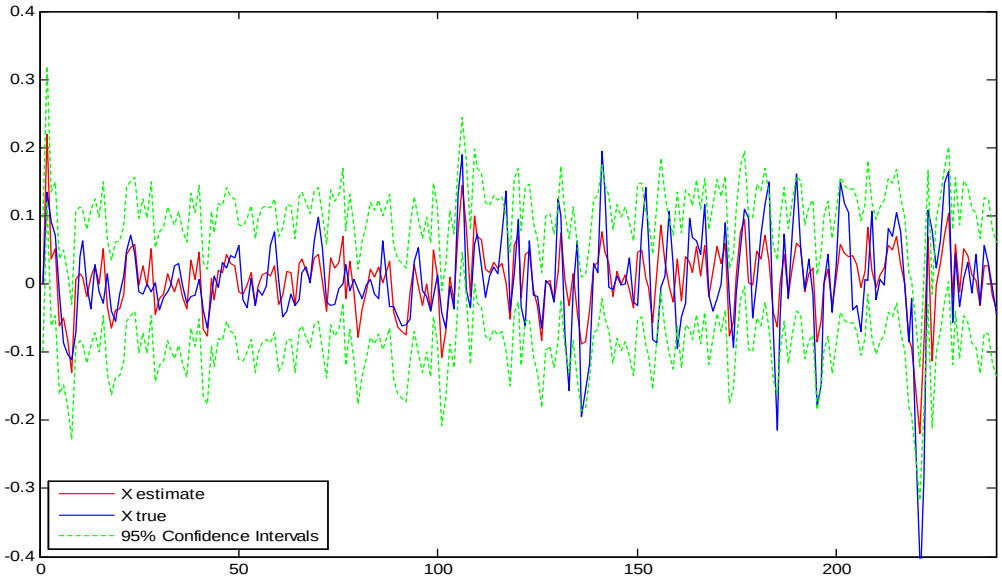

Figure 12: FEMA empirical application, 50 % omited, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

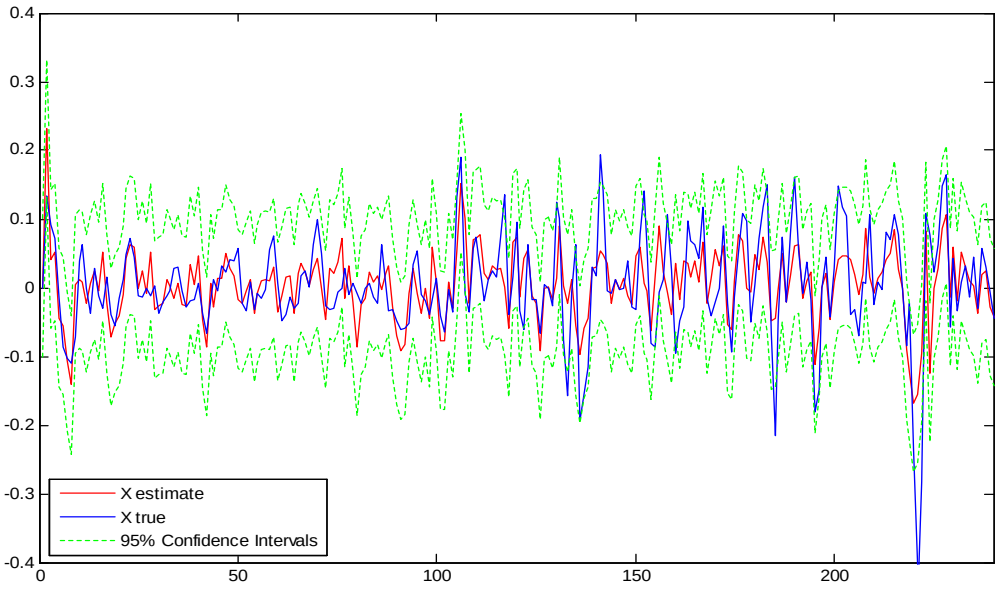

Figure 13: FEMA Empirical Application,75 % omitted, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

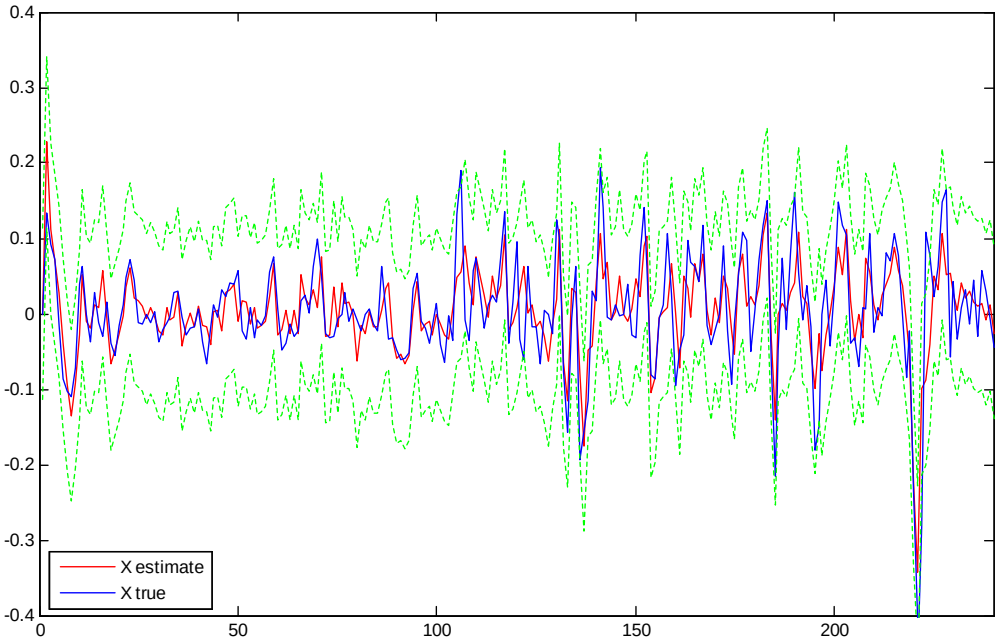

Figure 14: Factor spline Empirical Application,50 % omitted, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

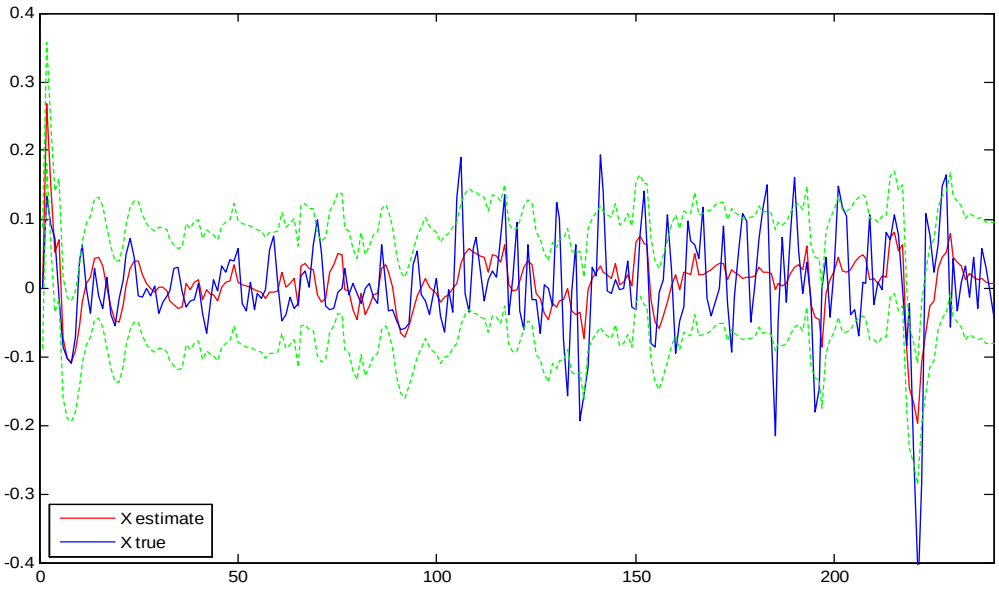

Figure 15: Factor spline Empirical Application,75 % omitted, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

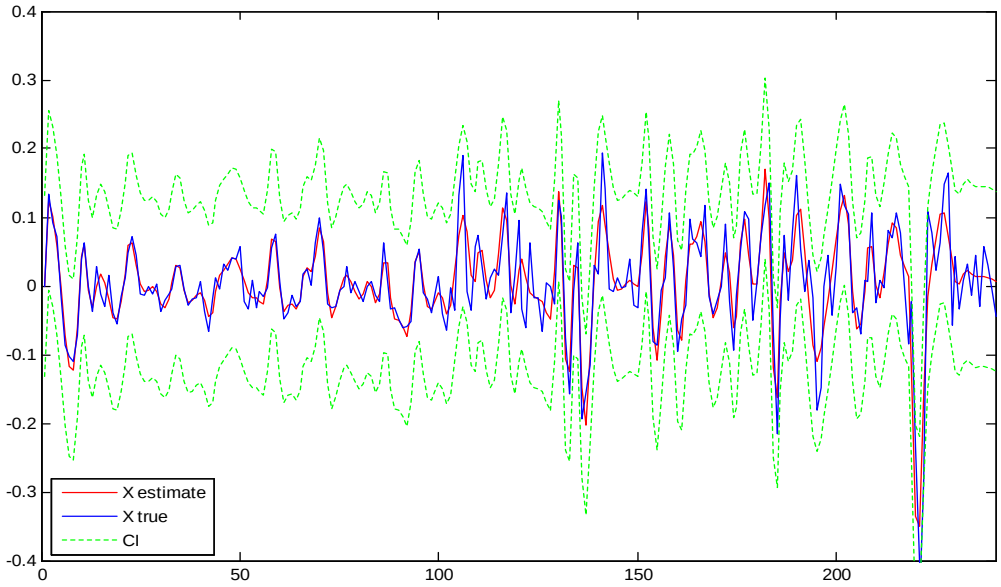

Figure 16: Spline Empirical Application,50 % omitted, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

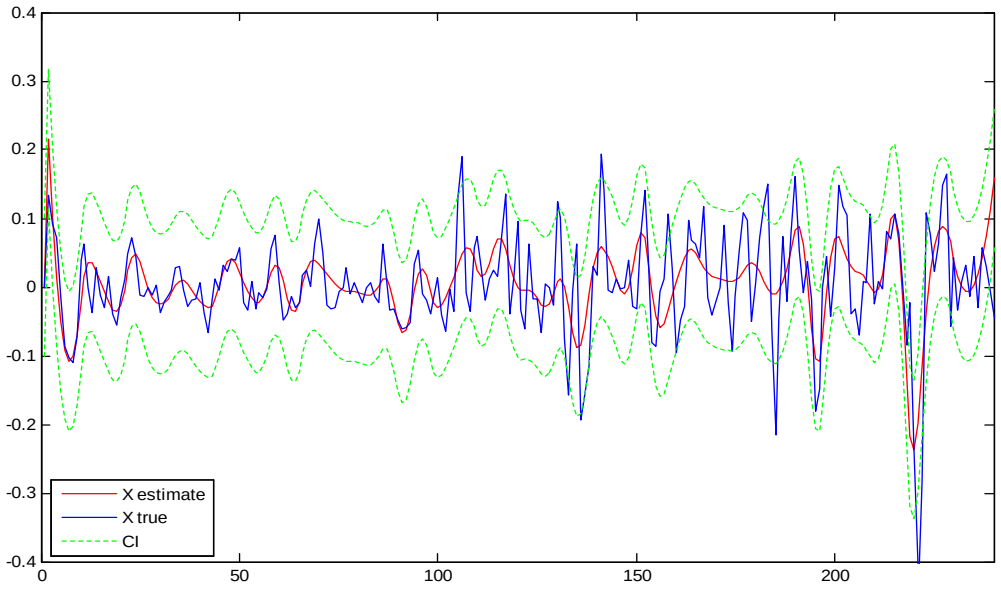

Figure 17: Spline Empirical Application,75 % omitted, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

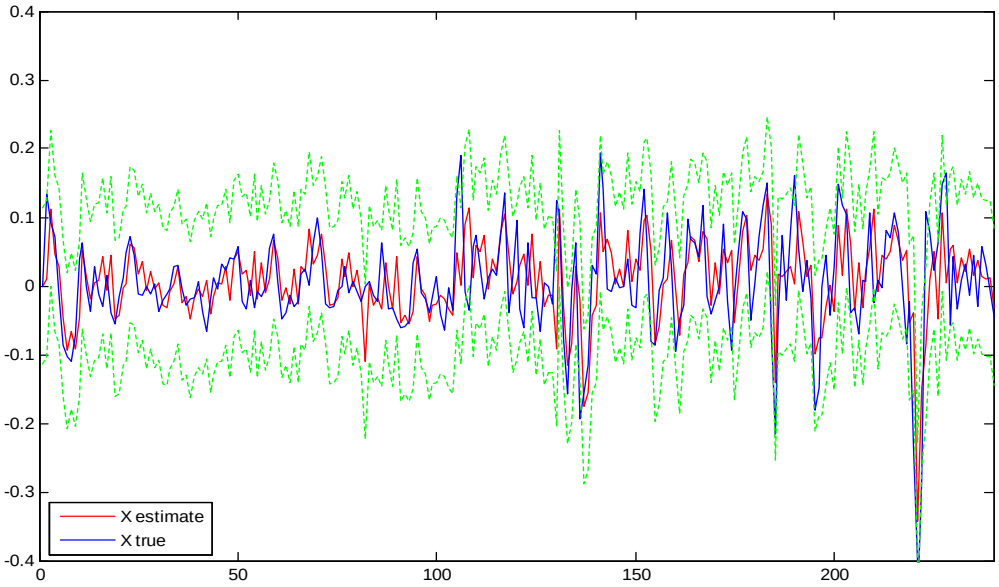

Figure 18: Factor Kalman Filter Empirical Application, 50 % omitted, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

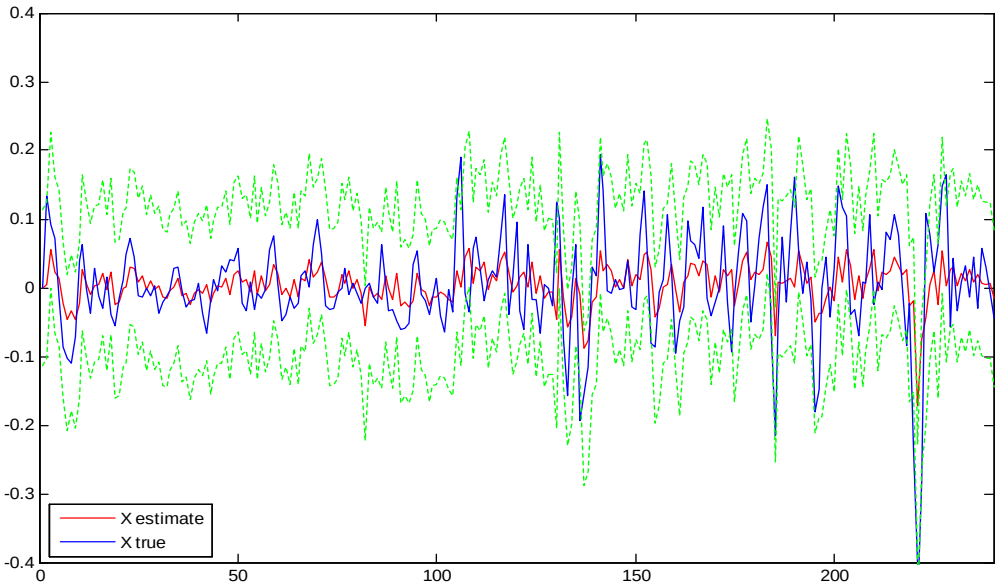

Figure 19: Factor Kalman Filter Empirical Application, 75 % omitted, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

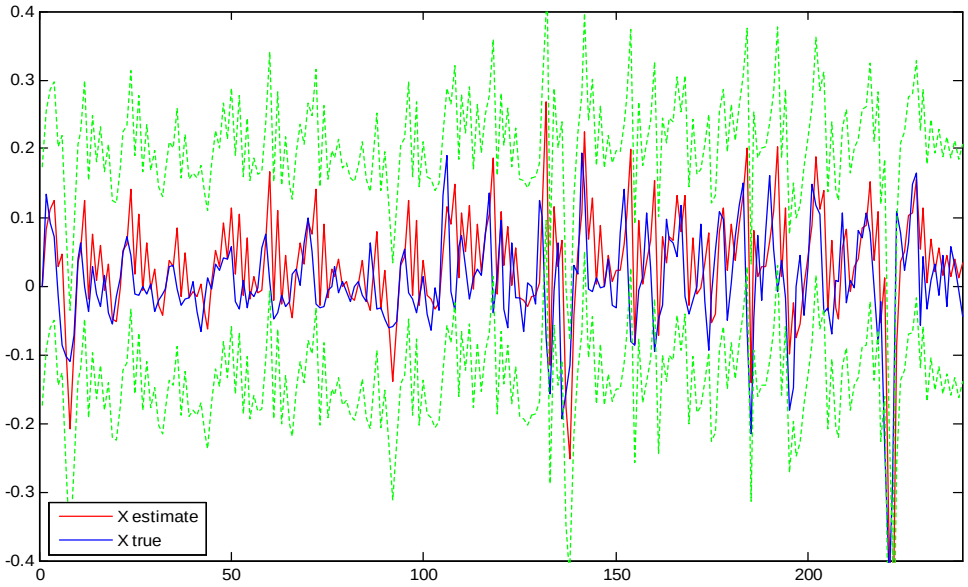

Figure 20: Kalman Filter Empirical Application, 50 % omitted, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

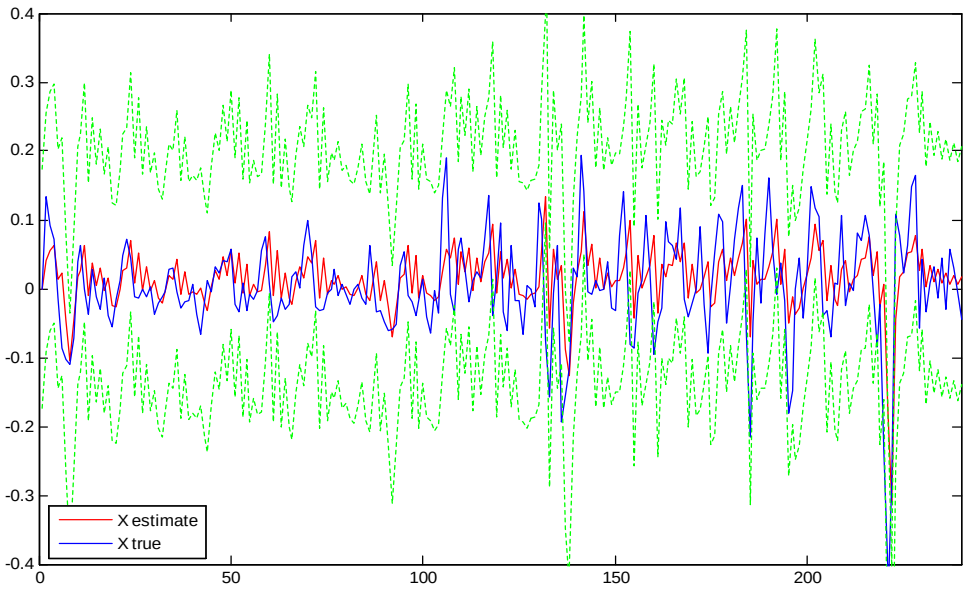

Figure 21: Kalman Filter Empirical Application, 75 % omitted, $X_{N,T} = (18, 240)$ $Y_{N,T} = (100, 240)$

Appendix D. List of refinery oil products

	Variable name	Measure	Source
1	Refiner Price of Finished Motor Gasoline to End Users	Dollars per Gallon	EIA
2	Refiner Price of Finished Aviation Gasoline to End Users	Dollars per Gallon	EIA
3	Refiner Price of Kerosene-Type Jet Fuel to End Users	Dollars per Gallon	EIA
4	Refiner Price of Kerosene to End Users	Dollars per Gallon	EIA
5	Refiner Price of No. 2 Fuel Oil to End Users	Dollars per Gallon	EIA
6	Refiner Price of No. 2 Diesel Fuel to End Users	Dollars per Gallon	EIA
7	Refiner Price of Propane (Consumer Grade) to End Users	Dollars per Gallon	EIA
8	Refiner Price of Finished Motor Gasoline for Resale	Dollars per Gallon	EIA
9	Refiner Price of Finished Aviation Gasoline for Resale	Dollars per Gallon	EIA
10	Refiner Price of Kerosene-Type Jet Fuel for Resale	Dollars per Gallon	EIA
11	Refiner Price of Kerosene for Resale	Dollars per Gallon	EIA
12	Refiner Price of No. 2 Fuel Oil for Resale	Dollars per Gallon	EIA
13	Refiner Price of No. 2 Diesel Fuel for Resale	Dollars per Gallon	EIA
14	Refiner Price of Propane (Consumer Grade) for Resale	Dollars per Gallon	EIA
15	Refiner Price of Residual Fuel Oil, Percent, Resale	Dollars per Gallon	EIA
16	Refiner Price of Residual Fuel Oil, Percent, End Users	Dollars per Gallon	EIA
17	Refiner Price of Residual Fuel Oil, Resale	Dollars per Gallon	EIA
18	Refiner Price of Residual Fuel Oil, End Users	Dollars per Gallon	EIA
19	Refiner Price of Residual Fuel Oil, Average, Resale	Dollars per Gallon	EIA
20	Refiner Price of Residual Fuel Oil, Average, End Users	Dollars per Gallon	EIA

Appendix E. List of variables in panel Y

	Variable name	Measure	Source
1	F.O.B. Cost of Crude Oil Imports From Mexico	Dollars per Barrel	EIA
2	F.O.B. Cost of Crude Oil Imports From Nigeria	Dollars per Barrel	EIA
3	F.O.B. Cost of Crude Oil Imports From Venezuela	Dollars per Barrel	EIA
4	F.O.B. Cost of Crude Oil Imports From Persian Gulf	Dollars per Barrel	EIA
5	Average F.O.B. Cost of Crude Oil Imports From All OPEC	Dollars per Barrel	EIA
6	Average F.O.B. Cost of Crude Oil Imports From All Non-OPEC	Dollars per Barrel	EIA
7	Landed Cost of Crude Oil Imports From Canada	Dollars per Barrel	EIA
8	Landed Cost of Crude Oil Imports From Mexico	Dollars per Barrel	EIA
9	Landed Cost of Crude Oil Imports From Nigeria	Dollars per Barrel	EIA
10	Landed Cost of Crude Oil Imports From Saudi Arabia	Dollars per Barrel	EIA
11	Landed Cost of Crude Oil Imports From Venezuela	Dollars per Barrel	EIA
12	Landed Cost of Crude Oil Imports From All OPEC	Dollars per Barrel	EIA
13	Landed Cost of Crude Oil Imports From All Non-OPEC	Dollars per Gallon	EIA
14	Unleaded Regular Gasoline, U.S. City Average Retail Price	Dollars per Gallon	EIA
15	Unleaded Premium Gasoline, U.S. City Average Retail Price	Dollars per Gallon	EIA
16	All Types of Gasoline, U.S. City Average Retail Price	Dollars per Gallon	EIA
17	Crude Oil and Natural Gas Rotary Rigs in Operation, Onshore	Number of Rigs	EIA
18	Crude Oil and Natural Gas Rotary Rigs in Operation, Offshore	Number of Rigs	EIA
19	Crude Oil Rotary Rigs in Operation	Number of Rigs	EIA
20	Natural Gas Rotary Rigs in Operation	Number of Rigs	EIA
21	Crude Oil and Natural Gas Rotary Rigs in Operation, Total	Number of Rigs	EIA
22	Active Well Service Rig Count	Number of Rigs	EIA
23	Wells Drilled, Exploratory, Crude Oil	Number of Wells	EIA
24	Wells Drilled, Exploratory, Natural Gas	Number of Wells	EIA
25	Wells Drilled, Exploratory, Dry	Number of Wells	EIA
26	Wells Drilled, Exploratory, Total	Number of Wells	EIA
27	Wells Drilled, Development, Crude Oil	Number of Wells	EIA
28	Wells Drilled, Development, Natural Gas	Number of Wells	EIA
29	Wells Drilled, Development, Dry	Number of Wells	EIA

30	Wells Drilled, Development, Total	Number of Wells	EIA
31	Wells Drilled, Total, Crude Oil	Number of Wells	EIA
32	Wells Drilled, Total, Natural Gas	Number of Wells	EIA
33	Wells Drilled, Total, Dry	Number of Wells	EIA
34	Crude Oil, Natural Gas, and Dry Wells Drilled, Total	Number of Wells	EIA
35	Total Footage Drilled	Thousand Feet	EIA
36	Asphalt and Road Oil Product Supplied	Thousand Barrels per Day	EIA
37	Aviation Gasoline Product Supplied	Thousand Barrels per Day	EIA
38	Distillate Fuel Oil Product Supplied	Thousand Barrels per Day	EIA
39	Jet Fuel Product Supplied	Thousand Barrels per Day	EIA
40	Propane/Propylene Product Supplied	Thousand Barrels per Day	EIA
41	Liquefied Petroleum Gases Product Supplied	Thousand Barrels per Day	EIA
42	Lubricants Product Supplied	Thousand Barrels per Day	EIA
43	Motor Gasoline Product Supplied	Thousand Barrels per Day	EIA
44	Petroleum Coke Product Supplied	Thousand Barrels per Day	EIA
45	Residual Fuel Oil Product Supplied	Thousand Barrels per Day	EIA
46	Total Petroleum Products Supplied	Thousand Barrels per Day	EIA
47	Crude Oil Imports, Total	Thousand Barrels per Day	EIA
48	Distillate Fuel Oil Imports	Thousand Barrels per Day	EIA
49	Jet Fuel Imports	Thousand Barrels per Day	EIA
50	Residual Fuel Oil Imports	Thousand Barrels per Day	EIA
51	Total Petroleum Imports	Thousand Barrels per Day	EIA
52	Crude Oil Stocks, SPR	Thousand Barrels	EIA
53	Crude Oil Stocks, Non-SPR	Thousand Barrels	EIA
54	Crude Oil Stocks, Total	Thousand Barrels	EIA
55	Distillate Fuel Oil Stocks	Thousand Barrels	EIA
56	Jet Fuel Stocks	Thousand Barrels	EIA
57	Motor Gasoline Stocks	Thousand Barrels	EIA
58	Residual Fuel Oil Stocks	Thousand Barrels	EIA
59	Crude Oil Refinery Net Input	Thousand Barrels	EIA

60	Natural Gas Plant Liquids Refinery and Blender Net Inputs	Thousand Barrels	EIA
61	Other Liquids Refinery and Blender Net Inputs	Thousand Barrels	EIA
62	Distillate Fuel Oil Refinery Net Production	Thousand Barrels	EIA
63	Jet Fuel Refinery Net Production	Thousand Barrels	EIA
64	Propane/Propylene Refinery Net Production	Thousand Barrels	EIA
65	Liquefied Petroleum Gases Refinery Net Production	Thousand Barrels	EIA
66	Finished Motor Gasoline Refinery and Blender Net Production	Thousand Barrels	EIA
67	Residual Fuel Oil Refinery Net Production	Thousand Barrels	EIA
68	Total Petroleum Refinery and Blender Net Production	Thousand Barrels	EIA
	MACROECONOMIC AND FINANCIAL DATA		
69	Yield on 10 year Gov US bonds percent Diff, log DS		Datastream
70	M1 billion dollars Diff, log DS		Datastream
71	M2 billion dollars Diff, log DS		Datastream
72	US Bank lending rate percent Diff, log DS		Datastream
73	US confidence index rate index Diff, log DS		Datastream
74	Producer's price index for finished goods index Diff, log DS		Datastream
75	US CPI index index Diff, log DS		Datastream
76	US industrial production index index Diff, log DS		Datastream
77	Yield on 20years US gov. Bonds percent Diff, log DS		Datastream
78	Sp500 index index Diff, log DS		Datastream
79	Yield on US 3yaers gov bonds percent Diff, log DS		Datastream
80	Share price of Exxon average price Diff, log DS		Datastream
81	Share price of BP average price Diff, log DS		Datastream
82	Share price of CONOCO average price Diff, log DS		Datastream
83	Share price of Shell average price Diff, log DS		Datastream
84	Share price of Chevron average price Diff, log DS		Datastream
85	Crude Oil-Dtd Brent UK Close USD/BBL price Diff, log DS		Datastream
86	Crude Oil-Brent 1Mth Fwd FOB USD/BBL price Diff, log DS		Datastream
87	US TREASURY BILL RATE - 3 MONTH (EP) percent Diff, log DS		Datastream
88	EURO to usd noon NY (EP) NADJ exchange rate Diff, log DS		Datastream

89	US-DS index OilGas - PRICE INDEX index Diff, log DS	Datastream
90	Citigroup Bond Index Corporate US index Diff, log DS	Datastream
91	Citigroup Bond Index Overall index Diff, log DS	Datastream
92	Citigroup Bond index treasury index Diff, log DS	Datastream
93	Citigroup Bond Index Industrial index Diff, log DS	Datastream
94	DAX stock market index index Diff, log DS	Datastream
95	UK stock market index index Diff, log DS	Datastream
96	China Industrial production index index Diff, log DS	Datastream
97	Euro area industrial production index index Diff, log DS	Datastream
98	USD-GBP exchange rate exchange rate Diff, log DS	Datastream
99	UK industrial production index index Diff, log DS	Datastream
100	World Dow-Jones industrial performance index Diff, log DS	Datastream

This page is intentionally left blank.

5 Superior predictive ability of data rich models: A study in oil futures

Ekaterina Ipatova

Cass Business School, City University London

April 9, 2014

Abstract:

A panel of common energy market fundamentals as well as macroeconomic variables is constructed in order to extract latent factors using the principle component methodology. The latent factors are then allowed to interact dynamically using a FAVAR (Factor Augmented Vector Autoregressive) model and FA-VECM (Factor Vector Error Correction) model which are used to develop short-term forecasts. A "*horse race*" of competing multivariate and univariate time series models is utilized in order to compare the forecasting performance of the factor augmented model. In order to mitigate the data snooping bias inherent in such studies we employ the non-parametric Hansen at el. (2011) MCS (Model Confidence Set) approach to evaluate the forecasting ability of the models, if any. We find that factor augmented models have superior sort term forecasting ability.

5.1 Introduction

Investment in commodities has seen a significant rise in the past decade with a myriad of indices and instruments presenting some lucrative opportunities. Similarly, a large proliferation of commodity based funds is also evident. A major component for investment attraction towards the commodity markets attributable to the lower correlation with other financial assets which provided diversification opportunities. Several prominent academic studies concluded that investors can significantly reduce portfolio risk and attract substantial risk premiums through relatively modest investments in long-only commodity index funds (e.g., Gorton and Rouwenhorst (2006); Erb and Harvey (2006)). Additional factors responsible for investors' embracing of the commodity markets is attributable to the general economic situation towards general economic situation during the middle of the past decade when interest rates declined to historically low levels, risk premiums steadily decreased, and traditional assets showed little obvious potential. Global outlook on the fast growing developing countries like Brazil, China and India, and the accompanying demand for oil, industrial metals and construction supplies determined positive market sentiment on the fundamental side of commodity investments and convinced investors in potential long-term high risk premiums (Kat and Oomen (2006)). The investment funds started to take advantage of deep and liquid exchange traded commodity futures, and more than 100 billion dollars moved in to commodity markets between 2004 and 2008. Domanski and Heath (2007) designate this as "financialization" of commodity futures markets. Investment boom reached its peak around the 2008 financial crisis, where the aggregate long positions held by commodity index investors reached about $256bn (see Mou and Yiqun (2010)).

Commodity financial instruments were introduced during the second half of the 20th century; however, the 2000s may be characterised by developments in the deep and liquid markets, such that additional benefits are felt as a result of a diverse asset portfolio(see, for example, Kat and Oomen et al. (2006)). The benefits are the result of a low correlation between commodity derivatives and traditional financial instruments (stocks and bonds). Aside the low correlation, investing in commodity markets was perceived to be a hedge against inflation as well. Browne and Cronin (2010) stated that traditional asset classes weaken and perform poorly during periods of sharply rising inflation. Commodities, on the other hand, benefit from rising inflation due to growth in price. There exists an opinion regarding endogeneity between commodity prices and inflation. Nevertheless, commodity derivatives help to protect a portfolio from the negative impact of inflation and systematic risks.

The benefits of commodity investments raise a question concerning optimal tools for the analyses and forecasting of the market. The primary motivation of this study is to provide a comparative analysis

of the predictive abilities of the factor base models in comparison to univariate models appropriate for the analysis of the commodity markets. A number of studies (e.g. Zagaglia (2010)) attempt to enhance predictive ability of the forecasting models on the energy markets using additional information extracted from large dimensional panels. The study focused on the application of the FA-VAR (Factor Vector Augmented methodology), where the dynamics of crude oil prices and common factors is modelled simultaneously in an attempt to increase forecastability of the oil markets. Our research posits a comparison of the FA-VAR model with the FA-VECM approach, which is applicable due to the co-integration between crude oil futures. Our hypothesis is that FA-VECM should produce superior results in comparison to FA-VAR, as it allows to model co-integrational relationships. We also provide the results for simple VAR and VECM models to demonstrate the benefits of the factor approach.

To order to use the factor approach in both models we incorporate more than 300 series, which are used in the factor panel. The panel dataset, therefore, incorporates supply and demand factors and also established financial and macro-economic variables. There is a downside to formulating a model based on large dataset which is that it is comprised of the series with mixed frequencies and data that is made available on different dates. This is known as a "ragged edge" problem (Ferrara et al 2010). A number of interpolation and smoothing algorithms are proposed (and used) to mitigate this. Following Bernanke (2008) and Zagaglia (2010) we extract common factors from a large dimensional panel and then model the joint dynamics of the 'latent' factors along with crude oil prices. We focus on FA-VECM (Factor Augmented Vector Error Correction Model) which tests superiority of the competing forecasts by adding latent trends from data-rich panels and additionally, incorporating information from term structure endogeneity (see Kilian(2008a) and Kilian (2008b)).

We propose a horse-race of competing models in order to determine the superiority of their predictive ability. Following the issue of data-snooping bias when a horse-race of competing models is carried out (see Hansen et al (2011), Stock and Watson (1999)), a possible solution may lie in formulating a loss function for in-sample performance and thereafter comparing it with an out-of-sample metric. However when we have multiple forecasts from competing models, ascertaining the superiority of one model over the benchmark is non-trivial. We therefore propose using the Model-Confidence Set (MULCOM) approach of Hansen et al (2011) which allows us to establish genuine outperformance and also to arrange the competing models according to their forecasting abilities.

This paper contributes to the literature by comparing the forecasting performance between data-rich models and univariate models. The superiority of the large dimensional models over the univariate ones

has never been tested in a robust framework. We use a non-parametric technique to test our hypothesis. We compare FA-VECM and a number of univariate ARFIMA-GARCH models

The remainder of the paper is organized as follows: section 5.2 describes the procedure we use to build two types of large dimensional based models: FAVAR, FAVECM. We provide the description of univariate models and, most importantly, model confidence set methodology, which is a non-parametric approach that determines model superiority. Section 5.3 describes the results of the model confidence set test, and section 5.4 summarise the paper and ends with some concluding remarks.

5.2 Methodology

A detailed description of the primary FAVAR and FA-VECM model is carried out along with the proposed augmentation of the procedure, as well as a description of competing models used to compare the outperformance of the primary series. In order to mitigate the data-snooping bias inherent to these studies, we describe the MULCOM procedure after Hansen (2011). A distinct advantage of using the MULCOM procedure over its previous counterparts' (the WRC & SPA) tests for data snooping is its ability to include nested formulations of a general model.

The methodology section is organised as follows: we start with a detailed description of the model and assumptions associated with the factor model. We give a complete account of all the procedures we use to establish existence and stability of the factor model used in the research. We continue with a description of the FA-VAR and FA-VECM model and describe the process of constructing the model. We continue with description of the competitive time-series models used as a comparison to the factor approach. We finish with a detailed description of the MULCOM procedure.

5.2.1 Factor Models

Latent Factors The foundations for the asymptotics and inferential theory of static factor models were laid down by Stock and Watson (2002a) and Bai (2003,2004).Following their contributions we establish a general form as well as a list of assumptions applied in the research. Formally, the static factor model is expressed as:

$$X_{it} = \lambda_i F_t + e_{it}$$

where X_{it} is the observation for $i'th$ cross section of the panel at time t; λ_i refers to factor loading, and F_t is a latent factor. We employ Principal Components (PC) methodology to convert panel datasets into the set of latent factors and loadings. Following Stock and Watson (2002a) we apply classical computational correction to the "short panels" where $T < N$. The estimated common factors, denoted by F_t, are therefore T times the eigenvectors corresponding to the largest eigenvalues of $T \times T$ matrix XX'. The transformation enables us to estimate common factors using an approach that is less computationally intensive. In comparison when we apply PC to long panels where $N < T$ we construct lambda first as the square root of N times eigenvectors corresponding to the largest eigenvalues of the $N \times N$ matrix $X'X$ and common factor computed using regression $F = (X \times \Lambda)/N$.

The application of PC to a large dimensional dataset allows us to use a general set of assumptions regarding error term e_{it} that is similar to the approach used in the current literature. Particularly, the error terms are not required to be normally distributed; there is a possibility of including weak cross-sectional dependence, weak time dependence, heteroskedasticity, weak dependence between factors and idiosyncratic errors (see Bai (2003)). Formally, a complete list of assumptions that are applicable to our model is detailed in section 2.3.1.

The methodology of our research aims at estimating a rolling FA_VAR and FA-VECM to obtain a series of one step-ahead forecasts which will be used to determine superior forecasting ability. During the estimation of rolling FA-VAR and FA-VECM and competing models we always work with the data inside the current rolling window r leaving the rest of the data for further windows. For consistency, we estimate the rolling factor model using an identical rolling window period (150 data points for all models) to select data from $T \times N$ dataset. In other words, we leave number of series N intact, but T for each estimation would be equal to r. The number of rolling periods is equal to $(T - r)$. As a result we obtain $r \times k \times (T - r)$ a matrix of common factors that are then used to estimate $(T - r)$ rolling factor models, where k is the optimal number of factors.

The primary objective of the rolling estimation in our case is to mitigate the possibility of any "forward-looking" bias that can potentially occur if we estimate matrix $T \times k$ common factors using the entire $T \times N$ dataset. Intuitively, the issue of the "forward-looking" bias is attributed to covariance used in PC. The means of variables used in covariances are very sensitive to the sample. If we include the sample $T \times N$ then the means of variables would be largely different from the means estimated from shorter rolling subsamples. In other words the means of two samples are not identical. The mean is a part of the covariance formula, and thus the covariance matrix will alter in accordance with the changes

in the mean. Factors are estimated using a covariance matrix and thus will also change. Moreover, means of variables estimated using $T \times N$ data include information for the entire period from $t = 1$ to T. At the same time while estimating rolling factor models we concentrate only on the data inside the current rolling window and the rest of the data is assumed unknown. Therefore, means of variables- if estimated using $T \times N$ dat,a bring information about the period assumed to be unknown for current window. Thus, the covariance matrix computed using $T \times N$ dataset passes this information to each common factor and it feeds information regarding future trends into the model in respect of current window. To avoid this "forward-looking" bias we estimate with the rolling factor model, so that each common factor reflects information only inside the matching rolling window.

We use the classical approach proposed by Bai and Ng (2002)[1] to determine the optimal number of factors. We estimate the number of common factors on the stationary panel, and identify that optimal number of factors equals 2. It is confirmed by estimation of the percentage of variation explained (PVE) by each factor for each of the rolling subsamples. PVE is equal to the eigenvalue corresponding to each factor divided by N. From Table XIX we can see that the first 2 factors explain the large percentage of variation. A drop of PVE for third and later factors explains the results of information criteria, which also demonstrate that the first two factors give the best approximation of common factor trends.

In the study we perform the Johansen co-integration test, to determine the presence of co-integration between 1, 3, 6 and 9 months futures of oil prices. The result is given in Table XXI. We can see that the co-integration test demonstrates 3 co-integration relationship. More importantly it signals the necessity of factor model, which take into consideration co-integration relationships.

Factor- Augmented VAR and VECM models A number of studies have used a factor framework that is formulated specifically to allow for endogeity (see Bernanke et al(2008), Zagaglia (2010)). In our study we decide to include restrictions within the model structure to allow for the classical factor model assumption, i.e. the factors are strictly orthogonal to one another and are also linearly independent. This will enable factor models to be more true to classical assumptions of factor models. In restricted models we permit endogeneity between observed components Y_t however we keep unobserved (latent) components F_t independent from each other. As a result unobserved factors are exogenous. The resulting FA-VAR and FA-VECM models are as follow:

[1] MatLab code for selection of optimal number of factors in factor model developed in Bai and Ng (2002) is available on: : http://www.columbia.edu/~sn2294/research.html.

$$
\begin{aligned}
Y_t &= \mu + \Phi(L)Y_{t-1} + \Pi(L)F_t + v_t \\
Y_t &= \mu + \Phi(L)Y_{t-1} + \Pi(L)F_t + \gamma u_{t-1} + v_t
\end{aligned}
$$

where $\Phi(L)$ and $\Pi(L)$ is a matrix of lag polynomials, and v_t is a vector of normally-distributed shocks, u_{t-1} is error correction term. Y_t is a vector of observed variables and F_t. is a vector of unobservable factors. Equation above states that the observabals are affected by each over, common factors and their own lags.

FA-VAR model is applicable to stationary series, therefore establishing stationarity of the observable Y_t. All elements of FA-VECM should be stationary according to the theory.In the "horse raise" using simple VAR and VECM models, whose equations are identical to the equation above with exception to the excluded factor augmentation part. Additionally we confirm stationarity of the exogenous part of the factor model: unobserved common factors. We perform Augmented Dickey-Fuller (see Dickey and Fuller (1979)) and Phillips-Perron (see Phillips and Perron (1988)) unit root test, where the number of lags for unit root tests is estimated using MAIC criteria proposed by Perron and Ng (2001)[2].

$$
MAIC(k) = \ln(\hat{\sigma}_k^2) + \frac{2(\tau_T(k) + k)}{T - k_{\max}}
$$

where $\tau_T(k) = (\hat{\sigma}_k^2)^{-1}\hat{\beta}_0^2 \sum_{t=k_{\max}+1}^{T} \tilde{y}_{t-1}^2$ and $\hat{\sigma}_k^2 = (T - k_{\max})^{-1} \sum_{t=k_{\max}+1}^{T} \hat{e}_{tk}^2$, and k is a lag order.

Augmented Dickey-Fuller (henceforth ADF) and Phillips-Perron (henceforth PP) unit root tests are estimated individually for all components of Y (oil future contracts) and F (common factors). Subsamples of Y and F tested separately for $(T-r)$ rolling iterations. The average test statistic across the individual rolling iterations is reported, for each of the future contracts as well as for the factors used in the model. According to the ADF and PP test, we find that the common factors are highly significant at the conventional ($\alpha = 0.05$) level of significance and are therefore stationary.

We change to AIC for optimal lag $\Phi(L)$ length selection of factor models where the maximum number of possible lags is eight. The optimal number of lags for all models across $(T - r)$ subsamples almost always equals one.

[2] Matlab codes for MAIC for the lag selection for unit root test is available on: http://people.bu.edu/perron/code.html.

To obtain the (static) one step ahead forecasts for the model, we use a rolling factor models regression with a window length of $r = 150$ observations. The choice of the window length is motivated by practical issues of model convergence. The window size is kept at 150 observations for all the other models for ease of comparison as well. The forecasting exercise consists of $(T - r) = 116$ iterations.

5.2.2 Time-Series Models

Research compares factor based models with the family of time-series models. We select the $ARFIMA-GARCH$ (Autoregressive Fractionally Integrated Moving Average) model to be the primary competitor to a factor based approach. We choose $ARFIMA-GARCH$ because it provides two distinct advantages. First it is parsimonious both in its structure and formulation and secondly it allows us to use multiple parameterizations. Therefore, we keep a unique base model, but increase the number of competing models by estimating $ARFIMA(p, d, q) - GARCH(\alpha, \beta)$ using different parameterisations of coefficients. The specification of the general model is as follows:

$$\begin{aligned} p(L)(1-L)^d Y_t &= q(L)e_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 e_{t-1}^2 + \beta\sigma_{t-1}^2 \end{aligned}$$

where Y_t is NYMEX crude oil futures, e_t is a serially uncorrelated, mean zero disturbance, $P(L) = 1 - p_1 L - ...p_h(L_h)$ is a stationary autoregressive process, and $q(L) = 1 + q_1 L + ...q_r L_r$ is an invertible moving-average process; σ_t^2conditional time-varying variance and ε_{t-1}^2 is realised volatility, d is fractional integration.

The mean of the process is modelled using Autoregressive Fractionally Integrated Moving Average, while conditional volatility is captured using $GARCH$. An issue with parameterising the conditional variance equation would be that the number of possible models (and their parameterisations) would be too numerous (see Poon and Granger (2003)). It has also been shown by Hansen, Nason and Lunde (2003) that amongst the various (330) $ARCH$ type models tested on a particular data set, there was very little evidence to suggest that there were superior formulations to the $GARCH(1, 1)$. However in order to capture the richer dynamics of the conditional volatility we increase parametrisation and as a result lags are selected from {1, 2, and 3}. The parameterizations for $ARFIMA$ part is obtained by varying

the combinations of the p and q lags between {0, 1, 2, 3}. After performing the estimation we obtain a total of 144 models those are all variations of $ARFIMA - GARCH$ model.

5.2.3 Model Confidence Set. Mitigating data-snooping bias

Data snooping was first referred to White (2000) in order to explain the issues arising out of re-using a given data set for the purposes of inference or possibly model selection. In our case we use the same panel and test a host of models over it. We are therefore not able to definitively conclude that any superior forecasting ability of the competing models is attributable chance alone. This is a typical issue inherent to time-series data where a single realization of the variable of interest is observable. There are various proposals laid out in the literature. The simplest would be an in-sample estimation period followed by an out-of-sample evaluation. Formal statistical tests include the Diebold and Mariano (1995) test and White's (2000) RC (Reality Check) procedure.

In order to mitigate the data snooping bias which is inherent in model forecast comparisons of this nature we can look to Hansen's (2005) methodology which is called the Superior Predictive Ability test (SPA) which allows us to test the null that there is no genuine forecasting outperformance of the best model chosen. We take our investigations a step further and decide to rank the models based on their outperformance by choosing the methodology of Hansen (2011) known as the Model Confidence Set. The advantage of the MCS over the SPA test is twofold, a) we do not have to choose a benchmark over which the other models are compared, and b) it returns the entire set of models which have superior performance by ranking them within a confidence set. In the spirit of earlier sequential testing procedures, White's (2000) RC for example; the procedure requires the following; a) an equivalence test, b) an elimination rule, and c) an updating algorithm.

Formally, M_0 denotes the set of (competing) forecasting models. M_0 consists of two subsets: M_{ts} which contain univariate time-series parameterizations and M_{FM} which contain variations of the factor based models in all $(T - r)$ individual forecasts. Each constituent model is indexed by $i \in \{1, ...m_0\}$. In order to illustrate the MCS procedure, we consider two competing forecast series $\{\hat{f}_{it,T}\}_{t=1}^{n}$ and $\{\hat{f}_{jt,T}\}_{t=1}^{n}$ with their corresponding forecast errors denoted by $\{\hat{e}_{t,T}^{i}\}_{t=1}^{n}$ and $\{\hat{e}_{t,T}^{j}\}_{t=1}^{n}$ generated by the $i'th$ and $j'th$ the competing models respectively. We specify MSE (Mean Squared Error) as a loss function for determining the forecasting ability of each competing model. This loss function could be easily replaced by a variety of other similar loss functions and is denoted by a general notation and $g(e_{t,T}^{i})$ and $g(e_{t,T}^{j})$.

This enables us to denote the set of superior (outperforming) models M^* as follows;

$$M* \equiv \{i \in M_0 : E(d_{t,T}^{ij}) \leq 0, j \in M_o\}$$

Where d_t^{ij} is the differential between the respective individual loss functions $g(.)$ at time t i.e. $d_t^{ij} = g(e_t^i) - g(e_t^j)$. At every iteration, based on $E\left(d_{t,T}^{ij}\right)$ one model is eliminated from the original set M_0 till M^* is reached. The criteria for iterative arriving is through the evaluation of the following null hypothesis:

$$H_0 : E(d_{t,T}^{ij}) = 0, \ i, j \in M$$

Upon rejection of the null, the candidate model j may be eliminated from the set M. Set M which reduces iteratively with the elimination of rules with poor forecasting ability, is said to have converged to the optimal set M^* when the null is accepted at a predefined level of significance α which is 0.05 in our case.

We follow Hansen (2011) in identifying the two primary components required to incorporate this testing procedure. They are the equivalence test and the elimination rule used in order to arrive at the confidence set denoted by $M_{1-\alpha}^*$. The equivalence test δ_M is based on the statistic $T_M \equiv \max_{i,j \in M} |t_{i,j}|$ where $t_{i,j} = \frac{\bar{d}_{ij}}{\sqrt{var(\bar{d}_{ij})}}$ and $\bar{d}_{ij} = \frac{1}{T} \sum_{t-1}^{T} d_{ij,t}$. The elimination rule states that the model selected for elimination is the one where $t = T_M$. Since the asymptotic distribution of the range is said to be non-standard according to Hansen (2011), they propose the use of the Politis and Romano (1994) block bootstrap.

5.3 Results/Data Analysis

5.3.1 The Dataset

Our study focusses on the energy markets from January 1990 to April 2012 for a total of $T = 266$ monthly observations. Vector of dependent variables Y_t includes future prices for WTI crude oil traded in the New York Mercantile Exchange (NYMEX) with date-to maturity $\{1, 3, 6, 12\}$ months. The one month (nearest futures contract) is used as a proxy for spot oil. The selection of contracts is justified by their liquidity (see Geman, Kharoubi (2008)). Also the specific selection is justified by our interest in the comparison

of modelling results between short term, medium term and long term contracts.

Panel data set comprised of a total number of $N = 301$ series that was originally utilised by Zagaglia (2010). In the research I updated original dataset and include 4 additional years which then cover the crises period. The dataset is well-matched for the current research for the number of reasons. First, factor based models require large dimensional datasets that would make factor based modelling possible. These requirements limit available series that can be included in the panel as some series may have a disproportionately low number of observations or series can be irrelevant distorting factor structure. Secondly, we believe that choice of variables in the dataset is entirely justified both on empirical as well as theoretical grounds. This is due to the fact that all series in the panel are meant to reflect macroeconomic, financial and geographical forces that drive oil prices, as well as number of tests proved empirically that dataset has a factor structure. Finally, a large proportion of the dataset has already been used for similar work and therefore shown to be relevant. The dataset is therefore both enhanced and updated we provide detailed justification/description of series below.

We start by including detailed statistics of energy resources consumed by individual sectors of the economy, as well as by the production sources, i.e renewable energy, natural gas and etc. To complete the world oil demand picture we include large number of series on petroleum consumption and storage for major OECD countries. A proxy of the consumption of crude oil is also included in the form of industrial production indices for emerging economies. We use industrial production indices as a higher frequency proxy of emerging countries GDP (as GDP available only on quarterly bases). GDP growth approximates the growth in production and therefore crude oil consumption (demand) pressure by emerging countries. The supply side of the oil market is described by including information about oil production in OPEC and non-OPEC countries. We include around 60 additional series that reflect information on drilling activity in US, costs of import from number of geographical regions and refinery prices. Special attention is paid to the US region which constitutes a third of the series within our panel. Close attention to the US region is due to the fact that our data shows that the US is a major consumer of energy products. In addition the US has a large publicly available dataset that describes energy market that goes back more than 20 years.

Our dataset is therefore constructed in a manner that reflects the fluctuations of major oil stocks and OTC derivative spreads reflecting speculative activity. It includes measures of monetary aggregates and indicators of confidence. US dollar exchange rates are included since it is argued that the stability of the US dollar is described as one of the key elements in identification of oil prices (see Zhang et al(2008)).

Major stock and bond indices are also included to provide a proxy for broader financial market sentiments prevailing at the time of the forecasts. Thus, our final dataset is a balanced panel consisting of monthly observations that are log differenced and de-trended to suit the requirements of the proposed primary (FAVAR) model. A complete list of the variables can be found in Appendix C.

5.3.2 Empirical Results

We apply two sets of competitive methodologies (factor models and univariate models) to the crude oil future contracts with short, medium and long maturities. We impose restrictions to build a set of 164 competing nested models. Factor models represented by FA-VECM and FA-VAR models, which accompanied by simple VECM and VAR models, 144 models obtained using ARFIMA-GARCH methodology and 16 models derived using simple ARFIMA. We use iterative procedure with rolling windows equal to 150 data points to extract one-step-ahead forecasts.

By gradually imposing a large number of restrictions we acquire a dataset where model one-step-ahead forecasts (and also five-step ahead forecast) changing gradually and minor from model to model. For Model Confidence Set (hence forth MCS) methodology the dataset with minor variations between forecasts is less informative than the one with wider variations (where difference between loss functions are larger). However, an attractive feature of the MCS approach is that it acknowledges the limitations of the data. Informative data will result in a MCS that contains only the best model. Less informative data make it difficult to distinguish between models and may result in a MCS that contains several (or possibly all) models. Thus, the MCS differs from extant model selection criteria that chooses a single model without regard to the information content of the data. In our research we expect more than one best (second best) performing model due to the fact that large portion of the dataset are nested models.

The complete dataset M0 contains 116 one-step-ahead forecasts for 164 models. We use M0 to start MCS procedure for identification of the superior predictive ability. The procedure performs 10000 bootstraps for 171 models with alpha=0.05 and p=0.15 with sensitivity check contained out for p=0.05 and 0.15 with no qualitative difference. As a result we acquire dataset M that is a ranking of superior predictive ability between models. According to our expectations M is large and contain several best (and second best) performing models.

The aim of the research is to explore the forecasting ability of data-rich panels. To do so we compare predictive ability of data-rich models with naïve univariate forecast. We are only interested in the top few

models that have superior predictive ability. If the factor model ranks in the top performing models we can insist that additional information that came from latent factors and term structure brings significant benefits. Thus we choose to report only top (30 models) forecasting models to be in line with objective of research and also to account for opportunity that more than one model can show the best results. The results are reported in Table XXII and XXIII.

[Insert Table XXII and XXIII around here]

In Table X we reported the results for crude oil futures contract with short (1 month), medium (3 and 6 months) and long (12 months) to maturities. The most important part of the results is p-value for the null hypothesis that the current model has superior predictive ability to the others. The higher p-value the better model's performance. P-values are estimated using the distribution of loss functions that are obtained by performing the bootstrapping procedure on individual forecasts for 164 models. Using this approach we are able to make statements about the significance of our results-a property that is not satisfied by the commonly used approach of reporting values from multiple pairwise comparisons. Along with p-values we report MSE value for each model, to be able to compare more robust MCS results to traditional loss function approach. Finally we report the rank of the model based on MCS p-value estimation.

The results vary across term structure of oil futures. We start from contract with 1 month to maturity. The best forecasting performance was shown by FAVECM model. We can observe that the second best result shown by ARFIMA-GARCH model and third place holds FAVAR model. Further we observe a large set of competing models which include VECM and VAR models. None of these competing models (p-value 0.7740) outperform each other when robust MSC methodology is used. We compare MSC results with classical MSE indicator. The results are consistent: FAVECM remains at the top with lowest MSE; it followed by ARFIMA(2,2)-GARCH(2,3) and FAVAR. Further, we observe wide dispersion between MSE that demonstrates the bias of pairwise comparison approach (out of sample comparison between forecast and real series).

The medium term crude oil futures contracts demonstrate different dynamics, with time–series models scoring higher p-value. For contracts with 3 months to maturity the set of factor models perform worse, with FAVECM being second best model and FAVAR being 14. Simple VECM and VAR perform worse than factor augmented models; for example, FAVECM holding second place and VECM ninth. Overall, in contract with 3 months to maturity we observed that factor model set is below top univariate models,

however FAVECM is very close to the top. It is interesting to note that while medium term factor models performed worse than time-series models, MSE results demonstrate that the loss function is decreased in comparison to the contract with 1 month to maturity. We interpret forecastability improvements by decreasing volatility across term structure (from closest towards furthest contracts). Factor models in medium term 6 months to maturity contracts set demonstrate worse forecastibility than in all compared options. FAVAR is not in the top 30 models, and FAVECM is only on 24 place; VECM is takes 29 place. At the same time MSE values variation is very narrow, indicating little difference between ability of the models to forecast.

Finally, 12 months to maturity contracts demonstrate exceptionally good results for factor model set. FAVECM and FAVAR takes first and third places and simple VAR and VECM second and forth best performance for long maturity contracts. Both factor models are included in the top set of predictive models signalling that latent factors consistently add superior predictive ability. The MSE value is the loss between the contracts that confirm the result that loss function decreases the value as we move across term structure.

Among 5-step ahead forecasts FAVECM is a leader model for 1,3 and 6 months to maturity contracts. However, the p-value demonstrate that ARFIMA(1,1) for 6-month contracts and FAVAR for 12 months contracts have equivalent predictive ability. FAVAR for 1,3 and 6 months contracts performs significantly worse than alternative factor model. Simple VECM and VAR perform worse than factor augmented models for all five-step ahead forecasts. The MSE values confirm the results of MCS as they gradually increasing from top best performing model to worst model.

Our results show plenty of useful guidelines. First we can see superiority of the factor augmented vector error correction models in majority for the forecast exercises. Additionally, we can observe that FAVECM demonstrates better results in comparison to FAVAR model, which proves that the information accumulated from co-integration relationship of crude oil prices with different maturities significantly increase accuracy of the forecast. At the same time worse performance of FAVAR can be due to misspecification of the model in comparison to FAVECM. Both FAVAR and FAVECM perform better than not augmented models (VECM and VAR) which proves that common factors increase forecasting ability of the models, which is similar to the previous findings of the related literature (see Zagaglia (2010)).

Factor models deliver consistently better results for forecasting entire term structures. Time-series models can be calibrated for better forecast of individual contracts but they does not allow to use

same univariate model specifications for the contracts with different maturities. We can conclude that application of the factor models and specifically FAVECM can stabilise forecastability of the model and performs superior forecasting for the entire term structure movements inside one regression.

5.4 Conclusion

We propose a "horse race" approach that juxtaposes the large dimensional and univariate models, and utilizes robust non- parametric procedures to determine superior predictive ability. Forecastability of crude oil future contracts has never been tested using robust MCS approach. Applying this methodology we obtain significant results contrary to those from a classical pairwise comparison. Using this technique we confirm the results by Zagaglia (2010) that factor models deliver consistent performance on the future oil markets. Our results are richer than we expect.

We can see that factor models are able to increase forecastability of the crude oil futures for the short and long term forecasting horizon. This is true across the term structure of crude oil futures. The best performing model is the factor augmented vector autoregressive model, which is able to accumulate additional information from the co-integrational relationship of crude oil futures and use it to improve the accuracy of the forecast. In comparison the factor augmented VAR model performs worse in comparison to FAVECM and many univariate models. If we compare factor augmented models with their non augmented counterparts we notice that factor augmentation can improve the performance of the model. Also factor models produce more stable results across the term structure; this is due to the fact that univariate models results vary across 4 analysed crude oil futures. Optimal univariate models for 1 month crude oil may not even be included in top forecasting models of the next crude oil futures. This issue can be observed across entire forecasting exercise.

We can conclude that we were able to prove superiority of the factor base model approach in the robust framework. Additionally, data-rich dataset improve the consistency of the forecast across the term structure. Therefore, we reach the goal of the research and reconfirm previous literature findings which were established using simple loss function (MSE). We can see further potential developments of the topic by investigating predictive ability of the data-rich models in comparison to more varied set of models.

Appendix A. List of Tables

Table XIX. Descriptive Statistics

Factor number	Average	Minimum	Maximum	.25 Quantile	Median	.75 Quantile
1	17.9651	14.8345	22.3217	15.6555	18.0321	19.9358
2	13.9687	12.1725	15.5392	13.8571	14.1371	14.3825

Table XIX demonstrates basic descriptive statistics for the distribution of factors used in the FA-VAR and FA-VECM models. This distribution is constructed from 116 factors estimated using a rolling regression.

Table XX. Unit-root average statistic accross rolling subsamples

Variable Name	1 m	3 m	6 m	12 m	F1	F2
PP	-11.9579	-11.2229	-10.5288	-10.0401	-23.6458	-10.0736
ADF	-8.2230	-7.5176	-9.3980	-9.9546	-8.8970	-8.7057

Table XX shows average statistics of Augmented Dickey-Fuller and Phillips Peron unit root tests for crude oil futures and factor trends. We fail to reject the null hypothesis that a process contains a unit root if the test statistic is larger than the critical values. *PP 5% critical value -1.9425. ** ADF 5% critical value -3.4419.

Table XXI Johancen Co-integration test

Unrestricted Cointegration Rank Test (Trace)			
No. of CE(s)	Trace Statistic	Critical Value 5%	Prob.**
None *	99.1580	47.8561	0.0000
At most 1 *	48.7564	29.7971	0.0001
At most 2 *	23.9919	15.4947	0.0021
At most 3	0.0602	3.8415	0.8061

Table XXI demonstrates the results of the Johansen co-integration test for crude oil future contracts with 1, 3, 6 and 12 months to maturity.

Table XXII. MCS Comparison of Crude Oil Future Contracts Forecast

Crude Oil Future Contract 1 M			Crude Oil Future Contract 3 M		
Model	MSE	p-value	Model	MSE	p-value
FA-VECM	0.0080	1.0000	ARFIMA(1,1)GARCH(3,2)	0.0072	1.0000
VECM	0.0081	0.9578	FA-VECM	0.0072	1.0000
FA-VAR	0.0087	0.7740	VECM	0.0073	0.9678
ARFIMA(2,2)GARCH(2,3)	0.0090	0.7740	ARFIMA(0,1)GARCH(2,1)	0.0073	0.9415
ARFIMA(2,2)GARCH(1,1)	0.0090	0.7740	ARFIMA(0,1)GARCH(3,2)	0.0073	0.9394
ARFIMA(1,0)	0.0090	0.7740	ARFIMA(1,0)GARCH(3,1)	0.0073	0.9068
ARFIMA(2,0)	0.0090	0.7740	ARFIMA(0,1)GARCH(3,1)	0.0073	0.8605
ARFIMA(0,0)GARCH(1,1)	0.0090	0.7740	ARFIMA(0,1)GARCH(2,3)	0.0074	0.8559
ARFIMA(3,1)GARCH(1,1)	0.0090	0.7740	ARFIMA(1,1)GARCH(3,1)	0.0074	0.7925
ARFIMA(1,0)GARCH(1,1)	0.0091	0.7740	ARFIMA(1,1)GARCH(2,2)	0.0074	0.7490
ARFIMA(0,1)GARCH(1,1)	0.0091	0.7740	ARFIMA(1,1)GARCH(2,1)	0.0074	0.7034
ARFIMA(0,1)	0.0091	0.7740	ARFIMA(0,1)GARCH(3,3)	0.0074	0.7060
ARFIMA(1,1)	0.0091	0.7740	ARFIMA(1,1)GARCH(3,3)	0.0074	0.6347
ARFIMA(2,1)	0.0091	0.7740	ARFIMA(1,0)	0.0075	0.5179
ARFIMA(0,3)	0.0091	0.7740	ARFIMA(2,0)	0.0075	0.5103
ARFIMA(0,0)	0.0092	0.7740	ARFIMA(0,1)	0.0076	0.5193
VAR	0.0092	0.7740	ARFIMA(2,1)	0.0077	0.5193
ARFIMA(3,1)	0.0092	0.7740	ARFIMA(0,2)	0.0077	0.5193
ARFIMA(0,2)	0.0092	0.7740	ARFIMA(1,2)	0.0077	0.5193
ARFIMA(1,2)	0.0092	0.7740	ARFIMA(0,0)GARCH(1,1)	0.0078	0.5193
ARFIMA(2,2)	0.0092	0.7740	ARFIMA(2,2)GARCH(1,1)	0.0078	0.5193
ARFIMA(1,3)	0.0092	0.7740	ARFIMA(0,0)	0.0078	0.5193
ARFIMA(3,2)GARCH(1,1)	0.0092	0.7740	ARFIMA(3,0)	0.0078	0.5193
ARFIMA(3,2)	0.0093	0.7740	ARFIMA(1,1)	0.0078	0.5193
ARFIMA(3,3)	0.0093	0.7740	FAVAR	0.0079	0.5193
ARFIMA(2,0)GARCH(1,1)	0.0093	0.7740	FAVAR	0.0079	0.5193
ARFIMA(1,1)GARCH(1,1)	0.0093	0.7740	ARFIMA(1,3)	0.0079	0.5193

Table XXII. *Continued.* MCS Comparison for Crude Oil Future Contracts Forecast

Crude Oil Future Contract 6 M			Crude Oil Future Contract 12 M		
Model	MSE	p-value	Model	MSE	p-value
ARFIMA(0,1)GARCH(3,2)	0.0061	1.0000	FAVECM with 2 factors	0.0045	1.0000
ARFIMA(1,0)	0.0062	0.8948	FAVECM with 3 factors	0.0045	1.0000
ARFIMA(1,0)GARCH(2,3)	0.0061	0.8938	ARFIMA(1,0)	0.0050	0.7950
ARFIMA(1,0)GARCH(3,2)	0.0061	0.8538	ARFIMA(2,0)	0.0051	0.7950
ARFIMA(0,1)GARCH(2,3)	0.0062	0.8538	ARFIMA(1,0)GARCH(1,1)	0.0051	0.7950
FAVECM	0.0062	0.8400	ARFIMA(0,1)GARCH(1,1)	0.0050	0.7950
VECM	0.0062	0.8460	ARFIMA(3,1)GARCH(1,1)	0.0051	0.7950
ARFIMA(1,0)GARCH(3,3)	0.0062	0.8460	ARFIMA(3,2)GARCH(1,1)	0.0051	0.7950
ARFIMA(2,0)	0.0063	0.8361	ARFIMA(0,1)GARCH(1,2)	0.0051	0.7950
ARFIMA(2,1)	0.0063	0.8361	ARFIMA(1,0)GARCH(2,1)	0.0051	0.7950
ARFIMA(1,2)	0.0064	0.8349	ARFIMA(0,1)GARCH(2,1)	0.0050	0.7950
ARFIMA(1,1)GARCH(1,2)	0.0063	0.8310	ARFIMA(3,2)GARCH(2,1)	0.0051	0.7950
ARFIMA(2,2)GARCH(1,2)	0.0063	0.8191	ARFIMA(1,0)GARCH(3,1)	0.0050	0.7950
ARFIMA(0,1)GARCH(1,3)	0.0063	0.8141	ARFIMA(0,1)GARCH(3,1)	0.0050	0.7950
ARFIMA(0,1)GARCH(2,2)	0.0062	0.8141	ARFIMA(0,1)GARCH(3,3)	0.0050	0.7950
ARFIMA(1,0)GARCH(3,1)	0.0063	0.8141	ARFIMA(1,1)	0.0051	0.7687
ARFIMA(0,1)GARCH(3,3)	0.0062	0.8101	FAVECM with 4 factors	0.0051	0.7604
ARFIMA(0,1)GARCH(3,3)	0.0062	0.8061	ARFIMA(1,0)GARCH(3,2)	0.0051	0.7484
ARFIMA(0,1)GARCH(1,2)	0.0063	0.8001	ARFIMA(0,1)GARCH(3,2)	0.0051	0.7274
ARFIMA(1,0)GARCH(1,2)	0.0063	0.8060	ARFIMA(0,1)GARCH(2,2)	0.0051	0.6986
ARFIMA(3,2)GARCH(3,1)	0.0065	0.7821	ARFIMA(0,1)GARCH(1,3)	0.0051	0.6113
ARFIMA(0,0)GARCH(3,2)	0.0064	0.7491	FAVECM with 1 factors	0.0051	0.6010
ARFIMA(0,1)GARCH(2,1)	0.0063	0.7471	FAVECM with 5 factors	0.0052	0.5969

Table XXII demonstrates the results of the "horse-race" between factor augmented models, univariate models and multivariate models. Table XXII presents the results evaluating the effectiveness of one-step ahead forecasts from different models using the loss function, MSE. Additionally, the table also reports the p-values obtained from the Hansen at el. (2011) MCS methodology. This compares the forecasting ability of the competing models by computing the MSE distribution using 10,000 bootstraps. The best forecasts will typically have a p-value close to or equal to 1, while the remaining p-values demonstrate

decreasing predictive ability of the models.

Table XXIII. MCS Comparison of Crude Oil Future Contracts Forecast. Multiple forecast

Crude Oil Future Contract 1 M			Crude Oil Future Contract 3 M		
Model	MSE	p-value	Model	MSE	p-value
FAVECM with 1 factor	0.0094	1.0000	FAVECM with 1 factor	0.0076	1.0000
ARFIMA(1,1)	0.0094	1.0000	FAVECM with 2 factors	0.0080	1.0000
ARFIMA(3,2)GARCH(1,1)	0.0101	0.9648	ARFIMA(1,1)GARCH(3,2)	0.0080	1.0000
ARFIMA(2,1)	0.0105	0.9457	ARFIMA(1,1)GARCH(3,3)	0.0106	0.9568
ARFIMA(2,3)	0.0109	0.9235	ARFIMA(3,2)GARCH(1,1)	0.0110	0.9457
ARFIMA(0,1)GARCH(1,1)	0.0111	0.9123	ARFIMA(0,1)GARCH(3,1)	0.0112	0.9457
ARFIMA(1,0)	0.0111	0.9123	ARFIMA(0,1)GARCH(2,3)	0.0112	0.9312
ARFIMA(2,0)	0.0111	0.9123	ARFIMA(1,1)GARCH(3,1)	0.0010	0.9312
ARFIMA(0,3)	0.0112	0.9123	ARFIMA(1,1)GARCH(2,2)	0.0111	0.9312
ARFIMA(3,2)	0.0115	0.9110	ARFIMA(1,1)GARCH(2,1)	0.0113	0.9312
ARFIMA(2,2)GARCH(2,3)	0.0131	0.9110	ARFIMA(0,1)GARCH(3,3)	0.0110	0.9312
FAVECM with 2 factors	0.0139	0.9110	ARFIMA(0,1)GARCH(2,1)	0.0119	0.8328
ARFIMA(2,2)GARCH(1,1)	0.0141	0.9110	ARFIMA(0,1)GARCH(3,2)	0.0119	0.8328
ARFIMA(3,1)GARCH(1,1)	0.0148	0.9110	FAVECM with 5 factors	0.0120	0.8328
FAVECM with 3 factors	0.0168	0.9110	FAVECM with 3 factors	0.0121	0.8328
ARFIMA(3,3)	0.0150	0.9110	FAVECM with 4 factors	0.0130	0.8328
ARFIMA(2,2)	0.0151	0.9110	ARFIMA(1,0)GARCH(3,1)	0.0120	0.8047
ARFIMA(0,0)GARCH(1,1)	0.0155	0.9110	ARFIMA(1,2)	0.0121	0.8047
ARFIMA(1,0)GARCH(1,1)	0.0160	0.9110	ARFIMA(2,2)GARCH(1,1)	0.0123	0.8047
ARFIMA(1,2)	0.0160	0.9110	ARFIMA(0,0)GARCH(1,1)	0.0125	0.8047
ARFIMA(2,0)GARCH(1,1)	0.0160	0.9110	ARFIMA(0,0)	0.0125	0.8047
ARFIMA(0,2)	0.0164	0.9110	ARFIMA(3,0)	0.0127	0.8047
ARFIMA(2,1)GARCH(1,1)	0.0166	0.9110	ARFIMA(1,1)	0.0128	0.8047
ARFIMA(3,0)	0.0167	0.9110	ARFIMA(1,0)	0.0128	0.8047
ARFIMA(0,1)	0.0169	0.9110	ARFIMA(2,0)	0.0128	0.8047
ARFIMA(1,1)GARCH(1,1)	0.0173	0.9110	ARFIMA(0,1)	0.0129	0.8047
ARFIMA(1,3)	0.0182	0.9110	ARFIMA(2,1)	0.0130	0.8047

Table XXIII. Continued MCS Comparison of Crude Oil Future Contracts Forecast. Multiple forecast

Crude Oil Future Contract 6 M			Crude Oil Future Contract 12 M		
Model	MSE	p-value	Model	MSE	p-value
ARFIMA(0,1)GARCH(3,2)	0.0121	1.0000	FAVECM with 2 factors	0.0125	1.0000
ARFIMA(1,0)	0.0123	0.9625	FAVECM with 3 factors	0.0125	1.0000
ARFIMA(1,0)GARCH(2,3)	0.0123	0.9625	ARFIMA(2,0)	0.0130	0.8012
FAVECM with 1 factor	0.0125	0.9537	ARFIMA(1,1)	0.0130	0.8012
FAVECM with 2 factors	0.0125	0.9537	ARFIMA(0,1)GARCH(1,2)	0.0130	0.8012
ARFIMA(1,0)GARCH(3,2)	0.0126	0.9537	ARFIMA(1,0)GARCH(2,1)	0.0131	0.8012
ARFIMA(0,1)GARCH(2,3)	0.0126	0.9537	ARFIMA(0,1)GARCH(2,1)	0.0131	0.8012
ARFIMA(1,1)GARCH(1,2)	0.0130	0.9234	ARFIMA(1,0)GARCH(1,1)	0.0135	0.7728
ARFIMA(1,2)	0.0131	0.9234	ARFIMA(1,0)GARCH(3,2)	0.0135	0.7728
ARFIMA(2,0)	0.0131	0.9234	FAVECM with 4 factors	0.0135	0.7728
ARFIMA(2,1)	0.0131	0.9234	ARFIMA(0,1)GARCH(1,1)	0.0136	0.7728
ARFIMA(2,2)GARCH(1,2)	0.0133	0.9234	ARFIMA(3,1)GARCH(1,1)	0.0136	0.7728
ARFIMA(0,1)GARCH(1,3)	0.0133	0.9234	ARFIMA(3,2)GARCH(1,1)	0.0136	0.7728
ARFIMA(1,0)GARCH(3,3)	0.0133	0.9234	ARFIMA(3,2)GARCH(2,1)	0.0136	0.7728
ARFIMA(1,0)GARCH(3,1)	0.0138	0.8930	ARFIMA(1,0)GARCH(3,1)	0.0137	0.7728
ARFIMA(0,1)GARCH(3,3)	0.0138	0.8930	ARFIMA(0,1)GARCH(3,1)	0.0137	0.7728
ARFIMA(0,1)GARCH(2,2)	0.0139	0.8930	ARFIMA(0,1)GARCH(3,3)	0.0137	0.7728
ARFIMA(0,1)GARCH(3,3)	0.0140	0.8930	ARFIMA(1,0)GARCH(2,2)	0.0138	0.7728
ARFIMA(0,1)GARCH(1,2)	0.0140	0.8930	ARFIMA(1,1)GARCH(2,2)	0.0139	0.7728
ARFIMA(1,0)GARCH(1,2)	0.0140	0.8930	FAVECM with 1 factors	0.0139	0.7728
ARFIMA(3,3)GARCH(3,2)	0.0142	0.8725	FAVECM with 5 factors	0.0140	0.7728
ARFIMA(2,2)	0.0142	0.8725	ARFIMA(0,1)GARCH(3,2)	0.0140	0.7728

Table XXII demonstrates the results of the "horse-race" between factor augmented models, univariate models and multivariate models. Table XXII presents the results evaluating the effectiveness of five-step ahead forecasts from different models using the loss function, MSE. Additionally, the table also reports the p-values obtained from the Hansen at el. (2011) MCS methodology. This compares the forecasting ability of the competing models by computing the MSE distribution using 10,000 bootstraps. The best forecasts will typically have a p-value close to or equal to 1, while the remaining p-values demonstrate decreasing predictive ability of the models.

Appendix B: List of variables in the panel

	Variable name	Measure	Source
1	F.O.B. Cost of Crude Oil Imports From Angola	Dollars per Barrel	EIA
2	F.O.B. Cost of Crude Oil Imports From Colombia	Dollars per Barrel	EIA
3	F.O.B. Cost of Crude Oil Imports From Mexico	Dollars per Barrel	EIA
4	F.O.B. Cost of Crude Oil Imports From Nigeria	Dollars per Barrel	EIA
5	F.O.B. Cost of Crude Oil Imports From Saudi Arabia	Dollars per Barrel	EIA
6	F.O.B. Cost of Crude Oil Imports From United Kingdom	Dollars per Barrel	EIA
7	F.O.B. Cost of Crude Oil Imports From Venezuela	Dollars per Barrel	EIA
8	F.O.B. Cost of Crude Oil Imports From Persian Gulf	Dollars per Barrel	EIA
9	Average F.O.B. Cost of Crude Oil Imports From All OPEC	Dollars per Barrel	EIA
10	Average F.O.B. Cost of Crude Oil Imports From All OPEC	Dollars per Barrel	EIA
11	Landed Cost of Crude Oil Imports From Angola	Dollars per Barrel	EIA
12	Landed Cost of Crude Oil Imports From Canada	Dollars per Barrel	EIA
13	Landed Cost of Crude Oil Imports From Colombia	Dollars per Barrel	EIA
14	Landed Cost of Crude Oil Imports From Mexico	Dollars per Barrel	EIA
15	Landed Cost of Crude Oil Imports From Nigeria	Dollars per Barrel	EIA
16	Landed Cost of Crude Oil Imports From Saudi Arabia	Dollars per Barrel	EIA
17	Landed Cost of Crude Oil Imports From Venezuela	Dollars per Barrel	EIA
18	Landed Cost of Crude Oil Imports From Persian Gulf	Dollars per Barrel	EIA
19	Landed Cost of Crude Oil Imports From All OPEC	Dollars per Barrel	EIA
20	Landed Cost of Crude Oil Imports From All Non-OPEC	Dollars per Barrel	EIA
21	Unleaded Regular Gasoline, U.S. City Average Retail Price	Dollars per Gallon	EIA
22	Unleaded Premium Gasoline, U.S. City Average Retail Price	Dollars per Gallon	EIA
23	All Types of Gasoline, U.S. City Average Retail Price	Dollars per Gallon	EIA
24	Refiner Price of Finished Motor Gasoline to End Users	Dollars per Gallon	EIA
25	Refiner Price of Finished Aviation Gasoline to End Users	Dollars per Gallon	EIA
26	Refiner Price of Kerosene-Type Jet Fuel to End Users	Dollars per Gallon	EIA
27	Refiner Price of Kerosene to End Users	Dollars per Gallon	EIA
28	Refiner Price of No. 2 Fuel Oil to End Users	Dollars per Gallon	EIA
29	Refiner Price of No. 2 Diesel Fuel to End Users	Dollars per Gallon	EIA

30	Refiner Price of Propane (Consumer Grade) to End Users	Dollars per Gallon	EIA
31	Refiner Price of Finished Motor Gasoline for Resale	Dollars per Gallon	EIA
32	Refiner Price of Finished Aviation Gasoline for Resale	Dollars per Gallon	EIA
33	Refiner Price of Kerosene-Type Jet Fuel for Resale	Dollars per Gallon	EIA
34	Refiner Price of Kerosene for Resale	Dollars per Gallon	EIA
35	Refiner Price of No. 2 Fuel Oil for Resale	Dollars per Gallon	EIA
36	Refiner Price of No. 2 Diesel Fuel for Resale	Dollars per Gallon	EIA
37	Refiner Price of Propane (Consumer Grade) for Resale	Dollars per Gallon	EIA
38	Refiner Price of Residual Fuel Oil, Percent, Resale	Dollars per Gallon	EIA
39	Refiner Price of Residual Fuel Oil, Percent, End Users	Dollars per Gallon	EIA
40	Refiner Price of Residual Fuel Oil, Resale	Dollars per Gallon	EIA
41	Refiner Price of Residual Fuel Oil, End Users	Dollars per Gallon	EIA
42	Refiner Price of Residual Fuel Oil, Average, Resale	Dollars per Gallon	EIA
43	Refiner Price of Residual Fuel Oil, Average, End Users	Dollars per Gallon	EIA
44	Coal Consumed by the Commercial Sector	Billion Btu	EIA
45	Natural Gas Consumed by the Commercial Sector	Billion Btu	EIA
46	Petroleum Consumed by the Commercial Sector	Billion Btu	EIA
47	Total Fossil Fuels Consumed by the Commercial Sector	Billion Btu	EIA
48	Hydroelectric Power Consumed by the Commercial Sector	Billion Btu	EIA
49	Geothermal Energy Consumed by the Commercial Sector	Billion Btu	EIA
50	Biomass Energy Consumed by the Commercial Sector	Billion Btu	EIA
51	Total Renewable Energy Consumed by the Commercial Sector	Billion Btu	EIA
52	Primary Energy Consumed by the Commercial Sector	Billion Btu	EIA
53	Electricity Retail Sales to the Commercial Sector	Billion Btu	EIA
54	Commercial Sector Electrical System Energy Losses	Billion Btu	EIA
55	Total Energy Consumed by the Commercial Sector	Billion Btu	EIA
56	Coal Consumed by the Electric Power Sector	Billion Btu	EIA
57	Natural Gas Consumed by the Electric Power Sector	Billion Btu	EIA
58	Petroleum Consumed by the Electric Power Sector	Billion Btu	EIA
59	Total Fossil Fuels Consumed by the Electric Power Sector	Billion Btu	EIA
60	Nuclear Electric Power Consumed by the Electric Power sector	Billion Btu	EIA
61	Hydroelectric Power Consumed by the Electric Power Sector	Billion Btu	EIA

94	Active Well Service Rig Count	Number of Rigs	EIA
95	Wells Drilled, Exploratory, Crude Oil	Number of Wells	EIA
96	Wells Drilled, Exploratory, Natural Gas	Number of Wells	EIA
97	Wells Drilled, Exploratory, Dry	Number of Wells	EIA
98	Wells Drilled, Exploratory, Total	Number of Wells	EIA
99	Wells Drilled, Development, Crude Oil	Number of Wells	EIA
100	Wells Drilled, Development, Natural Gas	Number of Wells	EIA
101	Wells Drilled, Development, Dry	Number of Wells	EIA
102	Wells Drilled, Development, Total	Number of Wells	EIA
103	Wells Drilled, Total, Crude Oil	Number of Wells	EIA
104	Wells Drilled, Total, Natural Gas	Number of Wells	EIA
105	Wells Drilled, Total, Dry	Number of Wells	EIA
106	Crude Oil, Natural Gas, and Dry Wells Drilled, Total	Number of Wells	EIA
107	Total Footage Drilled	Thousand Feet	EIA
108	Hydroelectric Power Consumed by the Electric Power Sector	Quadrillion Btu	EIA
109	Geothermal Energy Consumed by the Electric Power Sector	Quadrillion Btu	EIA
110	Solar/PV Energy Consumed by the Electric Power Sector	Quadrillion Btu	EIA
111	Wind Energy Consumed by the Electric Power Sector	Quadrillion Btu	EIA
112	Wood Energy Consumed by the Electric Power Sector	Quadrillion Btu	EIA
113	Waste Energy Consumed by the Electric Power Sector	Quadrillion Btu	EIA
114	Biomass Energy Consumed by the Electric Power Sector	Quadrillion Btu	EIA
115	Total Renewable Energy Consumed by the Electric Power Sector	Quadrillion Btu	EIA
116	Fuel Ethanol Feedstock	Trillion Btu	EIA
117	Fuel Ethanol Losses and Co-products	Trillion Btu	EIA
118	Fuel Ethanol Production	Trillion Btu	EIA
119	Fuel Ethanol Net Imports	Trillion Btu	EIA
120	Fuel Ethanol Stocks	Trillion Btu	EIA
121	Fuel Ethanol Consumption	Trillion Btu	EIA
122	Hydroelectric Power Consumed by the Industrial Sector	Quadrillion Btu	EIA
123	Geothermal Energy Consumed by the Industrial Sector	Quadrillion Btu	EIA
124	Fuel Ethanol Consumed by the Industrial Sector	Quadrillion Btu	EIA
125	Biomass Losses and Co-products in the Industrial Sector	Quadrillion Btu	EIA

126	Biomass Energy Consumed by the Industrial Sector	Quadrillion Btu	EIA
127	Total Renewable Energy Consumed by the Industrial Sector	Quadrillion Btu	EIA
128	Fuel Ethanol Consumed by the Transportation Sector	Quadrillion Btu	EIA
129	Biodiesel Consumed by the Transportation Sector	Quadrillion Btu	EIA
130	Biomass Energy Consumed by the Transportation Sector	Quadrillion Btu	EIA
131	Biofuels Production	Quadrillion Btu	EIA
132	Total Biomass Energy Production	Quadrillion Btu	EIA
133	Total Renewable Energy Production	Quadrillion Btu	EIA
134	Hydroelectric Power Consumption	Quadrillion Btu	EIA
135	Geothermal Energy Consumption	Quadrillion Btu	EIA
136	Solar/PV Energy Consumption	Quadrillion Btu	EIA
137	Wind Energy Consumption	Quadrillion Btu	EIA
138	Wood Energy Consumption	Quadrillion Btu	EIA
139	Waste Energy Consumption	Quadrillion Btu	EIA
140	Biofuels Consumption	Quadrillion Btu	EIA
141	Total Biomass Energy Consumption	Quadrillion Btu	EIA
142	Total Renewable Energy Consumption	Quadrillion Btu	EIA
143	Geothermal Energy Consumed by the Residential Sector	Quadrillion Btu	EIA
144	Solar/PV Energy Consumed by the Residential Sector	Quadrillion Btu	EIA
145	Wood Energy Consumed by the Residential Sector	Quadrillion Btu	EIA
146	Total Renewable Energy Consumed by the Residential Sector	Quadrillion Btu	EIA
147	Hydroelectric Power Consumed by the Commercial Sector	Quadrillion Btu	EIA
148	Geothermal Energy Consumed by the Commercial Sector	Quadrillion Btu	EIA
149	Wood Energy Consumed by the Commercial Sector	Quadrillion Btu	EIA
150	Waste Energy Consumed by the Commercial Sector	Quadrillion Btu	EIA
151	Fuel Ethanol Consumed by the Commercial Sector	Quadrillion Btu	EIA
152	Biomass Energy Consumed by the Commercial Sector	Quadrillion Btu	EIA
153	Total Renewable Energy Consumed by the Commercial Sector	Quadrillion Btu	EIA
154	Asphalt and Road Oil Product Supplied	Thousand Barrels	EIA
155	Aviation Gasoline Product Supplied	Thousand Barrels	EIA
156	Distillate Fuel Oil Product Supplied	Thousand Barrels	EIA
157	Jet Fuel Product Supplied	Thousand Barrels	EIA

158	Kerosene Product Supplied	Thousand Barrels	EIA
159	Propane/Propylene Product Supplied	Thousand Barrels	EIA
160	Liquefied Petroleum Gases Product Supplied	Thousand Barrels	EIA
161	Lubricants Product Supplied	Thousand Barrels	EIA
162	Motor Gasoline Product Supplied	Thousand Barrels	EIA
163	Petroleum Coke Product Supplied	Thousand Barrels	EIA
164	Residual Fuel Oil Product Supplied	Thousand Barrels	EIA
165	Other Petroleum Products Supplied	Thousand Barrels	EIA
166	Total Petroleum Products Supplied	Thousand Barrels	EIA
167	Crude Oil Imports, Total	Thousand Barrels	EIA
168	Distillate Fuel Oil Imports	Thousand Barrels	EIA
169	Jet Fuel Imports	Thousand Barrels	EIA
170	Propane/Propylene Imports	Thousand Barrels	EIA
171	Liquefied Petroleum Gases Imports	Thousand Barrels	EIA
172	Finished Motor Gasoline Imports	Thousand Barrels	EIA
173	Residual Fuel Oil Imports	Thousand Barrels	EIA
174	Other Petroleum Products Imports	Thousand Barrels	EIA
175	Total Petroleum Imports	Thousand Barrels	EIA
176	Crude Oil Exports	Thousand Barrels	EIA
177	Petroleum Products Exports	Thousand Barrels	EIA
178	Total Petroleum Exports	Thousand Barrels	EIA
179	Crude Oil Production, Persian Gulf	Thousand Barrels	EIA
180	Crude Oil Production, Canada	Thousand Barrels	EIA
181	Crude Oil Production, China	Thousand Barrels	EIA
182	Crude Oil Production, Egypt	Thousand Barrels	EIA
183	Crude Oil Production, Mexico	Thousand Barrels	EIA
184	Crude Oil Production, Norway	Thousand Barrels	EIA
185	Crude Oil Production, United Kingdom	Thousand Barrels	EIA
186	Crude Oil Production, United States	Thousand Barrels	EIA
187	Crude Oil Production, Total Non-OPEC	Thousand Barrels	EIA
188	Crude Oil Production, World	Thousand Barrels	EIA
189	Crude Oil Production, Algeria	Thousand Barrels	EIA

190	Crude Oil Production, Angola	Thousand Barrels	EIA
191	Crude Oil Production, Ecuado	Thousand Barrels	EIA
192	Crude Oil Production, Iran	Thousand Barrels	EIA
193	Crude Oil Production, Iraq	Thousand Barrels	EIA
194	Crude Oil Production, Kuwait	Thousand Barrels	EIA
195	Crude Oil Production, Libya	Thousand Barrels	EIA
196	Crude Oil Production, Nigeria	Thousand Barrels	EIA
197	Crude Oil Production, Qatar	Thousand Barrels	EIA
198	Crude Oil Production, Saudi Arabia	Thousand Barrels	EIA
199	Crude Oil Production, United Arab Emirates	Thousand Barrels	EIA
200	Crude Oil Production, Venezuela	Thousand Barrels	EIA
201	Crude Oil Production, OPEC	Thousand Barrels	EIA
202	Crude Oil Stocks, SPR	Thousand Barrels	EIA
203	Crude Oil Stocks, Non-SPR	Thousand Barrels	EIA
204	Crude Oil Stocks, Total	Thousand Barrels	EIA
205	Distillate Fuel Oil Stocks	Thousand Barrels	EIA
206	Jet Fuel Stocks	Thousand Barrels	EIA
207	Propane/Propylene Stocks	Thousand Barrels	EIA
208	Liquefied Petroleum Gases Stocks	Thousand Barrels	EIA
209	Motor Gasoline Stocks	Thousand Barrels	EIA
210	Residual Fuel Oil Stocks	Thousand Barrels	EIA
211	Residual Fuel Oil Stocks	Thousand Barrels	EIA
212	Residual Fuel Oil Stocks	Thousand Barrels	EIA
213	Crude Oil Refinery Net Input	Thousand Barrels	EIA
214	Natural Gas Plant Liquids Refinery and Blender Net Inputs	Thousand Barrels	EIA
215	Other Liquids Refinery and Blender Net Inputs	Thousand Barrels	EIA
216	Total Petroleum Refinery and Blender Net Inputs	Thousand Barrels	EIA
217	Distillate Fuel Oil Refinery Net Production	Thousand Barrels	EIA
218	Jet Fuel Refinery Net Production	Thousand Barrels	EIA
219	Propane/Propylene Refinery Net Production	Thousand Barrels	EIA
220	Liquefied Petroleum Gases Refinery Net Production	Thousand Barrels	EIA
221	Finished Motor Gasoline Refinery and Blender Net Producti	Thousand Barrels	EIA

222	Residual Fuel Oil Refinery Net Production	Thousand Barrels	EIA
223	Other Petroleum Products Refinery Net Production	Thousand Barrels	EIA
224	Total Petroleum Refinery and Blender Net Production	Thousand Barrels	EIA
225	Petroleum Consumption, France	Thousand Barrels	EIA
226	Petroleum Consumption, Germany	Thousand Barrels	EIA
227	Petroleum Consumption, Italy	Thousand Barrels	EIA
228	Petroleum Consumption, United Kingdom	Thousand Barrels	EIA
229	Petroleum Consumption, OECD Europe	Thousand Barrels	EIA
230	Petroleum Consumption, Canada	Thousand Barrels	EIA
231	Petroleum Consumption, Japan	Thousand Barrels	EIA
232	Petroleum Consumption, South Korea	Thousand Barrels	EIA
233	Petroleum Consumption, United States	Thousand Barrels	EIA
234	Petroleum Consumption, Other OECD	Thousand Barrels	EIA
235	Petroleum Consumption, Total OECD	Thousand Barrels	EIA
236	Petroleum Stocks, France	Million Barrels	EIA
237	Petroleum Stocks, Germany	Million Barrels	EIA
238	Petroleum Stocks, Italy	Million Barrels	EIA
239	Petroleum Stocks, United Kingdom	Million Barrels	EIA
240	Petroleum Stocks, OECD Europe	Million Barrels	EIA
241	Petroleum Stocks, Canada	Million Barrels	EIA
242	Petroleum Stocks, Japan Million Barrels EIA	Million Barrels	EIA
243	Petroleum Stocks, South Korea Million Barrels EIA	Million Barrels	EIA
244	Petroleum Stocks, United States Million Barrels EIA	Million Barrels	EIA
245	Petroleum Stocks, Other OECD Million Barrels EIA	Million Barrels	EIA
246	Petroleum Stocks, Total OECD Million Barrels EIA	Million Barrels	EIA
	Macroeconomic data		
247	Yield on 10 year Gov US bonds percent		Datastream
248	M1		Datastream
249	M2		Datastream
250	Capital utilization rate percentage index		Datastream
251	US confidence index rate index		Datastream
252	Producer's price index for finished goods index		Datastream

253	Producer's price index less food and energy index	Datastream
254	Federal Funds rate	Datastream
255	Consumption expenditure US	Datastream
256	US CPI index	Datastream
257	US industrial production index	Datastream
258	US house construction index	Datastream
259	Yield on 20years US gov. Bonds	Datastream
260	Dow Jones index	Datastream
261	Sp500 index index	Datastream
262	Yield on US 3yaers gov bonds	Datastream
263	Crude ligh 1 month open interest number of contracts	Datastream
264	Share price of Exxon average price	Datastream
265	Share price of BP average price	Datastream
266	Share price of CONOCO average price	Datastream
267	Share price of Shell average price	Datastream
268	Share price of Chevron average price	Datastream
269	JPMorgan global index	Datastream
270	JPMorgan global Eurobond index	Datastream
271	JPMorgna US gov bond index price	Datastream
272	Crude Spread WTI- Brent M+1 NY Cls price	Datastream
273	Crude Spread WTI- Brent M+2 NY Cls price	Datastream
274	Crude Spread Dubai M-M+1 NY Close price	Datastream
275	Crude Spread Dubai M+1-M+2 NY Close price	Datastream
276	Crude Oil-td Brent UK Close US	Datastream
277	Crude Oil-Brent 1Mth Fwd FOB US	Datastream
278	US TREASURY BILL RATE - 3 MONTH (EP) percent	Datastream
279	EURO to usd noon NY (EP) NA	Datastream
280	Morgan Stanley total index	Datastream
281	US-DS index Oil&Gas -Price Index	Datastream
282	Citigroup Bond Index Corporate US index	Datastream
283	Citigroup Bond Index Overall index	Datastream
284	Citigroup Bond index treasury index	Datastream

285	Citigroup bond Index Corporate Bond 1-3 years, Euro area	Datastream
286	Citigroup Bond Index Total Return index	Datastream
287	Citigroup Bond Index Industrial index	Datastream
288	Citigroup Bond Corporate Industrial Worldwide index	Datastream
289	DAX stock market index	Datastream
290	UK stock market index	Datastream
291	China Industrial production index	Datastream
292	Euro area industrial production index	Datastream
293	US/GBP exchange rate exchange rate	Datastream
294	UK industrial production index	Datastream
295	World Dow-Jones industrial performance index	Datastream
296	CBOE VIX (implied volatility index) index	Datastream
297	NYMEX Natural gas 2 month price	Datastream
298	NYMEX Natural gas 3 month price	Datastream
299	NYMEX Natural gas 6 month price	Datastream
300	NYMEX Heating oil 2 month price	Datastream
301	NYMEX Heating oil 3 month price	Datastream

This page is intentionally left blank.

This page is intentionally left blank.

6 Conclusion

This thesis addressed the current issues in the field of factor models. We concentrated on three distinct problems contributing to the theoretical, methodological and empirical literature related to factor models theory. We developed our research from the bases of evaluating the existing literature on factor models theory and identifying gaps in the literature. As a result we recognized three topics, the development of which would be beneficial to the expansion of the existing factor models literature.

In the first chapter, we proposed a novel methodology for estimating common factors, factor loadings and common components from the data in levels for first difference models. Alternative methodology has same first order terms in comparison to the existing estimators, but offer different asymptotic results for the higher order terms. As a result, estimators computed according to our novel methodology have higher rate of convergence and give more robust approximation of the model parameters. The aim of our research is to demonstrate a novel methodology aiming to enhance the rates of convergence for the estimated factor trends and common components, without additional assumptions.

The research extends the tradition of the modern factor model literature, which over the past fifty years moved towards relaxation of the basic assumptions, improvement of the consistency of the estimators and removing the boundaries of the panel dataset. Original assumptions imposed strict restrictiond on the dataset dimensions, and prevented the occurrence of heteroskedustisity, autocorrelation, and cross-sectional correlation of the error terms in the factor model. Additionally, the errors had to be normally distributed. Further developments help to relax these assumptions, whilst ensuring convergence of the estimators to the true factor. The assumption of classical large dimensional literature (Bai (2003), Bai (2004), and Stock and Watson (2002)) demonstrate the convergence of the parameters, however, our methodology helps to improve rates of convergence in comparison to the classical literature. It implies that with using similar assumptions we are able to achieve more robust and consistent factor trends.

Monte-Carlo simulations were performed to test the methodology. Factors were simulated using an autoregressive process; loadings were drawn from univariate distributions. The error terms of factor models were simulated using multiple specifications of the process that can be described using the ARMA model. We find that the majority of the common factors estimated using the novel methodology has a marginally higher degree of correlation with the true simulated factor, in comparison to the factors estimated using the classical approach. We also found that the degree of correlation between the true factors and estimated factors increased as the dimensions of the panel increased. We found that the estimated

factor converges to the true factor given the large dataset, and both novel and classical methodologies lead to the similar results but only for the exceptionally large panels. Smaller panels demonstrate a marginal difference between the correlation of true factor and estimated factor, with a higher correlation resulting from the novel methodology. Our research helps to improve the consistency of the factor estimator, which has wide theoretical and empirical application. From a theoretical prospective the research adds to the long line of theoretical findings which aimed to improve estimation methodology, without loss of generality. From an empirical prospective, our methodology allows to estimate factors which have higher rates of convergence to true factors and therefore improve the quality of the estimators. We are of the opinion that research can be further developed and greater rates of convergence achieved. More interesting still will be investigation of the rates of convergence in the relatively medium and small panels, as we observe a larger impact of the novel methodology on these types of panels. Additionally, further development of the topic should progress towards greater relaxation of the basic assumptions of the factor models, without loss of the consistent and robust performance of the estimators.

The third chapter of the research thesis is focused on the identification and filling of the missing observations in the factor panel. The motivation for the research comes from the fact that current factor model literature predominantly focused on the estimation of the factors, and analysing the factor based forecast. However, to perform factor analyses research has to construct a balanced panel of data which will allow application of the principle components method and estimation of factors. Modern literature has paid little attention to the topic of the construction of the appropriate factor models, especially when it comes to the area of filling missing observations in the unbalanced panels. The majority of the literature (see literature review in chapter 3) focuses on the individual missing observations in the panel, or alternatively on the missing observations at the beginning or the end of the dataset. Our methodology provides the alternative, which allows one to fill observations in any part of the panel, i.e. at the end or beginning of the dataset, as well as blocks of missing observations, and individual missing observations in the panel. The methodology attempts to employ this feature of the factor model, using factor structure to solve the problem of missing observations simultaneously. It allows one to map any pattern of missing observations and re-construct the observations, given that number of assumptions are fulfilled. The most significant limitation of our methodology is that the researcher has to be able to extract the alternative common factor which does not have any missing observations, nor a high degree of correlation with the factor from the unbalanced panel. We argue that factor datasets can be divided into balanced and unbalanced panels, where the first panel contains all variables with missing observations, and the second panel contains all the balanced variables. By definition, factor panel should share common trend(s)

between all variables in the panel, and therefore factors from the unbalanced and balanced panels should share a common trend. This assumption is applicable to empirical research, as it is common that some parts of the panel contain well balanced variables, while other variables have some missing observations.

We perform both the simulation and empirical exercises in order to test the validity of our methodology. We simulate panels with different degrees of distortion, such that the panels can have between 50% and 80% missing observations. Additionally, we performed this exercise for the panels with a different number of observations and variables. We compare the technique with four alternative models. Two competing models contained the factor base approach; we used the factor panels to estimate common factors and later employed Kalman Filter and cubic spline to nowcast missing observations directly in the factor. Later, we used the factor with filled omitted observations to restore the panel dataset using the process of matrix multiplication of the factor and factor loading. Two alternative approaches use Kalman filter and cubic splyne technique to fill missing observations directly in the factor panel. We report goodness of fit of the nowcasting procedure by reversing Theil's U loss function, which is estimated on the basis of the normalised RMSE loss function. The results are bound between 0% and 100%, with higher results indicating better fit of the missing observations.

We report average goodness of fit of the fitted values, as well as a variation of the goodness of fit of the missing observations from a change in the panel size. We can see that our methodology demonstrates the highest and most stable results in filling missing observations. Theil's U function demonstrates an average result of 58%, which is stable if we increase the number of missing observations to 80%. This is explained by the methodology, which provides an approximation of the true factor fluctuations using a closely correlated alternative common factor. Spline methodology performs significantly worse. This is due to the fact that for the large blocks of missing observations, splyne tends to converge to the factor mean and is therefore unable to demonstrate any fluctuations during that period. The Kalman filter need some stable preliminary blocks of data where the parameters can be learned. The Kalman filter will perform significantly better for the missing observations at the end of the forecasted period, however in the panels with a mixed pattern of missing observations, the kalman filter does not have enough data to calibrate the parameter. We also performed an investigation of the structural parameters. We found that structural parameters are different for high and low frequency estimations. The difference between structural parameters is due to variations in the sample size; while size of T dimension in the 100% balanced panel is 240 observations; the sample size in the panel with 80% missing observations is 48. Such a drastic difference in the sample sizes of the regression leads to the difference in the structural

parameters.

We can conclude that the methodology developed in the research is able to fill missing observations with greater accuracy than competing models, which is shown by range of sensitivity tests. We provide theoretical proofs to support our research methodology. We would like to continue work on further developing the methodology. At the moment we can see a few potential areas: firstly, we believe that the range of competing models can be expanded; secondly, concurrently we only provide approximation to the common factor using the second highly correlated factor. We never investigate a possibility of the factor loading approximation, and instead impose the reasonable assumption of the stability of the factor loading. Additionally, the research can be extended to provide greater examination of the appropriate selection of the factor used for approximation in the model.

The final chapter of the thesis focuses on the problem of empirical application of the factor model theory. We use energy markets to illustrate the superior forecasting ability of the factor model. The superiority of the factor model builds on the notion that factor models are able to filter the noise from large dimensional panels and extract the common trend between all the variables in the panel. It is impossible to include all the variables of large dimensional panels in the regression due to the fact that we easily run short in our degree of freedom. In addition, the computational intensity is vast in regressions where the number of repressors is assumed to be infinite, and such a regression will thus no longer prove informative. The number of parameters also has to be reduced, such that we can use all relevant information to improve the forecast.

We use two factor augmented models FA-VAR and FA-VECM. We take into consideration co-integration between crude oil markets in the FA-VECM model; both models are compared with their non-factor augmented multivariate VAR and VECM models, which use only crude oil returns to forecast future fluctuations. Additionally, we use a set of univariate models ARFIMA-GARCH which attempt to model crude oil returns only by elaborating on the dynamic of the time-series. We propose a "horse race" approach that juxtaposes the large dimensional and univariate models, and utilizes robust non-parametric procedures to determine superior predictive ability. We use Hansen's et al. (2011) approach to establish superior forecasting; the approach employs the bootstrap technique and estimates the best forecasting model using 10,000 bootstrap iterations.

According to the estimations, the FA-VECM model demonstrates the best performance amongst competing factor models for WTI crude Oil futures forecasts with 1 month, 3 months, 6 months and

12 months to maturity. We can observe that FA-VECM leads among the competing univariate and multivariate model forecasts. We can conclude that co-integration relationship play a significant role in the forecast of the crude oil market. At the same time, factor variables also improve the model performance in comparison to the non-factor augmented models across all the one-step and multi-step forecasts. Therefore, our research helps to fill the gap in the empirical literature of factor models by examining forecasting ability of factor based models in a robust way, using Hansen's et al. (2011) methodology for the first time. We also examine the FA-VECM model for forecasting in the crude oil market and achieve superior results. These are the two major contributions of the final chapter of the thesis. In the future the research may be extended to include more compatible models in the "horse race" of Hansen's et al. (2011) methodology, which should help to establish new ways to improve the forecasting of the crude oil market.

This page is intentionally left blank.

Appendix. Asymptotic theory

The appendix presents additional assumptions, lemmas, propositions and theorems which applied in the asymptotic proofs for sections 3 and 4. We concentrate on the assumptions that are required for the proofs. Assumptions and conditions for proofs of the lemmas and propositions, but not directly used in the current study proofs are not present. They can be found in the original works. All proofs in the study follow the original assumptions and conditions. We start from assumptions and Lemmas applied to the non-stationary large dimensional panels, which are discussed in original work by Bai(2004):

- *Assumption IFor each* i, as $T \to \infty$

$\frac{1}{T}\sum_{t=1}^{T} F_t^0 e_{it} \xrightarrow{d} \int B_u dB_e^{(i)}$

Assumption I is following Bai(2004) p148, that defines B_u is $r \times 1$ vector of Brownian motions defined in section 2.3.1; $B_e^{(i)}$ is scalar Brownian motion process with variance $\Omega_{ee}^{(i)} = \lim(1/T)\sum_{t=1}^{T}\sum_{s=1}^{T} E(e_{it}e_{is}).B_u$ and $\mathrm{B}_e^{(i)}$ are independent. We apply assumption I to obtain limiting distribution of correlation between two factors in the third section.

Our proofs use set of lemmas from Bai(2004) those are presented below:

- *Lemmas Bai(2004):*

(i) Lemma A.1 p164 Under assumptions A-C, we have for some $M < \infty$, and for all N and T.

$E\left(T^{-1}\sum_{t=1}^{T}\left\|N^{-1/2}e_t'\Lambda^0\right\|^2\right) = E\left(T^{-1}\sum_{t=1}^{T}\left\|N^{-1/2}\sum_{i=1}^{N} e_{it}\lambda_i^0\right\|^2\right) \leq M,$

$E\left(T^{-4}\sum_{t=1}^{T}\sum_{s=1}^{T}\left(N^{-1}\sum_{i=1}^{N} X_{it}X_{is}\right)^2\right) \leq M,$

$E\left\|(NT)^{-1/2}\sum_{i=1}^{N}\sum_{t=1}^{T} e_{it}\lambda_i^0\right\| \leq M.$

(ii) Lemma B.1 p 167 Under assumptions A-D, we have

$\delta_{NT}^2 \left(\frac{1}{T} \sum_{t=1}^{T} \left\| \breve{F}_t - H_2' F_t^0 \right\|^2 \right) = O_p(1)$

where $H_2 = H_1 V_{NT}^{-1}$ has full rank and V_{NT} is an $r \times r$ diagonal matrix consisting of the first r largest eigenvalues of $(1/NT^2)XX'$ in decreasing order; $\delta_{NT} = \min\left\{\sqrt{N}, T\right\}$.

(iii) Lemma B.4(i) p171 Under assumptions A-E form Bai(2004), the $r \times r$ matrix sutisfies:

$T^{-1}(\hat{F} - F^0 H_1)' F^0 = O_p(T^{-1}) + O_p(N^{-1/2})$,

We also consider a *proposition 1 Bai(2004) p 143* in the proof for section 3:

$\max_{1 \le t \le T} \left\| \hat{F}_t^k - H^{k\prime} F_t^0 \right\| = O_p(T^{-1}) + O_p(\sqrt{T/N})$

- *Theorem 2 p148 Bai (2004): As* $N, T \to \infty$, with $N/T^3 \to 0$, we have for each t

$\sqrt{N}\left(\hat{F}_t - H' F_t^0\right) = \left(\frac{\tilde{F}' F}{T^2}\right) \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda_i^0 e_{it} + O_p(1)$

$\xrightarrow{d} QN(0, \Gamma_t)$

where $\tilde{F}' F^0 / T^2 \underrightarrow{d} Q$ and $\Gamma_t = \lim_{n\to\infty} (1/N) \sum_{i=1}^{N} \sum_{j=1}^{N} E(\lambda_i^0 \lambda_j^{0\prime} e_{it} e_{jt})$

- *Theorem 3 p149, Bai (2004) As* $N, T \to \infty$,

$T(\hat{\lambda}_i - H_1^{-1} \lambda_i^0) = H_1^{-1} \left(\frac{F^{0\prime} F^0}{T^2}\right)^{-1} \sum_{t=1}^{T} F_t^0 e_{it} + O_p(1)$

$\xrightarrow{d} (\Sigma_\Lambda Q')^{-1} (\int B_u B_u') \int B_u dB_e^{(i)}$,

where Σ_Λ is positive definite non-random matrix and $\tilde{F}' F^0 / T^2 \xrightarrow{d} Q$

- *Proposition 1 Bai(2004) p.143*

$\max_{1 \le t \le T} \left\| \hat{F}_t^k - H^{k\prime} F_t^0 \right\| = O_p(T^{-1}) + O_p(\sqrt{T/N})$

Assumptions,Lemmas formulated in Bai(2003):

- *Assumption J Bai(2003) p144* for each i, as $T \to \infty$,

$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} F_t^0 e_{it} \xrightarrow{d} N(0, \Phi_i)$,

where Φ_i =plim$_{T\to\infty}(1/T) \sum_{s=1}^{T} \sum_{t=1}^{T} E\left[F_t^0 F_s^{0\prime} e_{is} e_{it}\right]$

- *Lemmas Bai(2003):*

(i) Lemma A1 p158

$\tilde{F} - H'F^0 = V_{NT}^{-1}(\frac{1}{T}\sum_{s=1}^{T} \tilde{F}_s \gamma_N(s,t) + \frac{1}{T}\sum_{s=1}^{T} \tilde{F}_s \zeta_{st} + \frac{1}{T}\sum_{s=1}^{T} \tilde{F}_s \eta_{st} + \frac{1}{T}\sum_{s=1}^{T} \tilde{F}_s \xi_{st})$

where $\zeta_{st} = e_s' e_t/N - \gamma_N(s,t)$ and $\eta_{st} = F_s^{0\prime}\lambda^{0'} e_t/N$ and $\xi_{st} = F_t^{0\prime}\lambda^{0\prime} e_s/N$ and V_{NT} is a diagonal matrix consisting of the first r eigenvalues of $(1/NT)XX'$ in decreasing order.

(ii) Lemma B2 p164 Under

$T^{-1}(\tilde{F} - F^0 H)' F^0 = O_p(\delta_{NT}^{-2})$

where $\delta_{NT}^2 = \min\{N, T\}$

- *Theorem 1 p145 Bai(2003)* As $N, T \to \infty$

(i) if $\sqrt{N}/T \to 0$, then for each t

$\sqrt{N}(\tilde{F}_t - H F_t^0) = V_{NT}^{-1}\left(\frac{\tilde{F}' F^0}{T}\right)\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \lambda_i^0 e_{it} + O_p(1) \xrightarrow{d} N(0, V^{-1} Q \Gamma_t Q' V^{-1})$,

where V_{NT} is given in section 6; matrix Q is invertible and is given by $Q = V^{1/2}\Upsilon'\Sigma^{-1/2}$, where $V = diag(v_1, v_2, ..., v_r), v_1 > v_2 ... > v_r > 0$ are eigenvalues of $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$, and Υ is the corresponding eigenvector matrix such that $\Upsilon'\Upsilon = I_r$;and $\Gamma_t = \lim_{N\to\infty}(1/N)\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i^0\lambda_j^{0\prime} E(e_{it} ejt)$.

(ii)if $\liminf \sqrt{N}/T \geq \tau > 0$, then

$T(\tilde{F}_t - H F_t^0) = O_p(1)$.

- *Theorem 2 p147, Bai(2003)* As $N, T \to \infty$

(i) if $\sqrt{T}/N \to 0$, then for each i,

$\sqrt{T}(\bar{\lambda}_i - H^{-1}\lambda_i^0) = V_{NT}^{-1}\left(\frac{\tilde{F}'F^0}{T}\right)\left(\frac{\Lambda^{0\prime}\Lambda^0}{N}\right)\frac{1}{\sqrt{T}}\sum_{t=1}^{T} F_t^0 e_{it} + O_p(1) \xrightarrow{d} N(0, (Q')^{-1}\Phi_i Q^{-1})$,

where V_{NT} and Φ_i are defined in section 6; matrix Q is invertible and is given by $Q = V^{1/2}\Upsilon'\Sigma^{-1/2}$, where $V = diag(v_1, v_2, ..., v_r), v_1 > v_2 ... > v_r > 0$ are eigenvalues of $\Sigma_\Lambda^{1/2}\Sigma_F\Sigma_\Lambda^{1/2}$, and Υ is the corresponding eigenvector matrix such that $\Upsilon'\Upsilon = I_r$;

(ii) if $\liminf \sqrt{N}/T \geq \tau > 0$, then

$N(\tilde{\lambda}_i - H^{-1}\lambda_i^0) = O_p(1)$

The dominant case is part (i), asymptotic normality. Part (ii) is of theoretical interest.

- *Trapani (2012a) Theorem 1. p.130 , for* $(N, T) \to \infty$ and for all i

$\frac{1}{T^2}\sum_{t=1}^{T} F_t F_t' \xrightarrow{d} H'(\int W_\epsilon W_\epsilon')H$,

$\frac{1}{T}\sum_{t=1}^{T} F_t u_{it} \xrightarrow{d} H' W_\epsilon dW_{u,i}$

where W_ϵ is a k-dimensional Brownian motion with covariance matrix $\Sigma_{\Delta F}$, and $W_{u,i}$ is a scalar Brownian motion independent of W_ϵ.

This page is intentionally left blank.

References

[1] Altissimo, F. (2001) "A real time coincident indicator of the euro area business cycle", Centre for Economic Policy Research;

[2] Anderson, T. W., Rubin, H. (1956) "Statistical inference in factor analysis",. In: J. Neyman (ed.): Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. V. Berkeley: University of California Press, pp. 114–150;

[3] Anderson, T. W. (1958) "An introduction to multivariate statistical analysis", Wiley New York;

[4] Angelini, E., Jérôme H., Marcellino, M. (2006) "Interpolation and backdating with a large information set", Journal of Economic Dynamics and Control: 2693-2724;

[5] Armah, N. A., Swanson, N. R. (2010) "Seeing inside the black box: Using diffusion index methodology to construct factor proxies in large scale macroeconomic time series environments", Econometric Reviews: 476-510;

[6] Armah, N. A., Swanson, N. R. (2011) "Some variables are more worthy than others: new diffusion index evidence on the monitoring of key economic indicators", Applied Financial Economics : 43-60;

[7] Artis, M., Banerjee, A., Marcellino, M. (2005) "Factor forecasts for the UK", Journal of Forecasting 24, 279–298;

[8] Baffes, J., Haniotos, T. (2010) "Placing the 2006/08 Commodity Boom into Perspective", Policy Research Working Paper 5371, The World Bank, Washington, D.C.;

[9] Bai, J., Ghysels, E., Wright, J. H. (2009) "State space models and MIDAS regressions", Econometric Reviews;

[10] Bai, J., Ng, S. (2002) "Determining the number of factors in approximate factor models", Econometrica 70(1), 191–221;

[11] Bai, J., Ng, S. (2006a) "Confidence intervals for diffusion index forecasts and inference with factor-augmented regressions", Econometrica 74(4), 1133–1150;

[12] Bai, J., Ng, S. (2011) "Principal components estimation and identification of the factors", manuscript, Columbia University ;

[13] Bai, J. (2003) "Inferential theory for structural models of large dimensions", Econometrica. 71, 135-171;

[14] Bai, J. (2004) "Estimating cross-section common stochastic trends in nonstationary panel data", Journal of Econometrics, 122(1), 137-183;

[15] Bai, J., Ng, S. (2004) "A PANIC attack on unit roots and cointegration", Econometrica, 72(4), 1127-1177;

[16] Bai, J., Ng, S. (2002) "Determining the number of factors in approximate factor models", Econometrica 70.1 191-221;

[17] Bai, J., Ng, S. (2006) "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions", Econometrica 74.4 : 1133-1150;

[18] Baillie, R., Bollerslev, T., Mikkelsen, H. O. (2006) "Fractionally integrated generalized autoregressive conditional heteroskedasticity" Journal of Econometrics, 74 -1, 3-30;

[19] Bandalos, D. L., Boehm-Kaufman, M. R. (2009) "Four common misconceptions in exploratory factor analysis", Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences, 61-87;

[20] Bartholomew, D. J., Knott, M., Moustaki, I. (2011) "Latent variable models and factor analysis: a unified approach", (Vol. 899). Wiley;

[21] Bartlett, M. S. (1938) "The approximate recovery of information from replicated field experiments with large blocks", J. Agric. Sci, 28, 418-427;

[22] Bernanke, B. S., Boivin, J. (2003) "Monetary policy in a data-rich environment", Journal of Monetary Economics 50.3: 525-546;

[23] Bernanke, B. S., Bovin, J., Eliasz, P. (2008) "Measuring the effect of monetary policy: a Factor Augumented Vector Autoregressive (FAVAR) approach",Quarterly Journal of Economics 120 (1) 387-422;

[24] Bessembinder, H., Seguin, P. J. (1993) "Price volatility, trading volume, and market depth: Evidence from futures markets", Journal of financial and Quantitative Analysis 28.1;

[25] Biernacki, C., Celeux,G., Govaert,G. (2003) "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models", Computational Statistics & Data Analysis 41.3 : 561-575;

[26] Boivin, J., Ng, S. (2005) "Understanding and comparing factor-based forecasts", (No. w11285) National Bureau of Economic Research;

[27] Brennan, M. (1958) "The Supply of Storage", American Economic Review 48: 50-72;

[28] British Petroleum Statistical Review (2012), available at: www.bp.com/statisticalreview;

[29] Browne, F., Cronin, D. (2010) "Commodity Prices, Money and Inflation", Journal of Economics and Business;

[30] Camacho, M., Pérez-Quirós, G., Poncela, P. (2012) "Markov-switching dynamic factor models in real time", Available online: http://www.eea-esem.com/files/papers/eea-esem/2011/664/cpqpp_all.pdf;

[31] Chamberlain, G., Rothschild, M. (1983) "Arbitrage, factor structure and mean-variance analysis in large asset markets", Econometrica 51(5), 1281–1304;

[32] Chow, G. C., Lin, A. (1971) "Best linear unbiased interpolation, distribution, and extrapolation of time series by related series", The review of Economics and Statistics: 372-375;

[33] Connor, G., Korajczyk, R. A. (1986) "Performance measurement with the arbitrage pricing theory: A new framework for analysis", Journal of financial economics, 15(3), 373-394;

[34] Connor, G., Korajczyk, R. A. (1988) "Risk and return in an equilibrium APT: Application of a new test methodology", Journal of Financial Economics, 21(2), 255-289;

[35] Connor, G., Korajczyk, R.A. (1993), "A Test for the Number of Factors in an Approximate Factor Model", Journal of Finance, XLVIII, 1263-1291.Analysis in Large Asset Markets," Econometrica, 51, 1305-1324;

[36] Considine, T. J., Larson, D. F. (2001) "Uncertainty and the convenience yield in crude oil price backwardations", Energy economics, vol23, issue5, 533–548;

[37] Cootner, P. (1967) "Stock prices: random vs. systematic changes", Industrial Management Review, 3 (1962), pp. 24–45;

[38] De Boor, C. (1978). A practical guide to splines (Vol. 27). New York: Springer-Verlag.

[39] De Mol, C., Giannone, D., Reichlin, L. (2008) "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?", Journal of Econometrics, 146(2), 318-328;

[40] De Roon, F. A., Nijman, T. E., Veld, C. (2000) "Hedging pressure effects in futures markets", The Journal of Finance 55.3, 1437-1456;

[41] De Schutter, O. (2010) "Food Commodities Speculation and Food Price Crises: Regulation to Reduce the Risks of Price Volatility", Briefing Note 02 by the United Nations Special Rapporteur on the Right to Food;

[42] Dickey, D. A., Fuller, W. A. (1979) "Distributioonf Estimator for AutoregressivTime Series With a Unit Root", Journaol of theAmerican Statistical Association, 74, 427-431;

[43] Diebold, F. X., Mariano, R. S. (1995) "Predictive Accuracy", Jornal of Business ad Economic Statistic 13 , 253-263;

[44] Dincerler, E., Cantekin, A., Khoker, Z., Titman, S. (2003) "Futures Premia and Storage", University of Western Ontario, working paper;

[45] Diron, M. (2008) "Short-term forecasts of euro area real GDP growth: an assessment of real-time performance based on vintage data", Journal of Forecasting 27.5: 371-390;

[46] Domanski, D., Heath, A. (2007) "Financial Investors and Commodity Markets", Bank for International Settlements Quarterly Review (March), p: 53–67;

[47] Dusak, K. (1973) "Futures Trading and Investor Returns: An Investigation of Commodity Market Risk Premiums.", J.P.E. 81, 1387-1406;

[48] Eickmeier, S., Ziegler, C. (2008) "How successful are dynamic factor models at forecasting output and inflation? A meta-analytic approach", Journal of Forecasting 27.3 : 237-265;

[49] Engle, R. (2002) "Dynamic conditional correlation", Journal of Business & Economic Statistics, 20(3), 339-350;

[50] Erb, C. B., Campbell R. H. (2006) "The Strategic and Tactical Value of Commodity Futures", Financial Analysts Journal 62(2): 69–97;

[51] Evans, M. D. (2005) "Where are we now? Real-time estimates of the macro economy", (No. w11064). National Bureau of Economic Research;

[52] Fama, E., French, K. (1987) "Commodity Futures Prices: Some Evidence on Forecast Power, Premiums, and the Theory of Storage", Journal of Business 60: 55-73;

[53] Fama, E., French, K. (1988) "Business Cycles and the Behaviour of Metals Prices", Journal of Finance 43: 1075-1093;

[54] Favero, C., Marcellino, M., Neglia, F. (2005) "Principal components at work: The empirical analysis of monetary policy with large datasets", Journal of Applied Econometrics 20, 603–620;

[55] Ferrara, L., Gugan, D., Rakotomarolahy, P. (2010) "GDP nowcasting with ragged-edge data: a semi-parametric modeling", Journal of Forecasting, 29 (1-2), 186–199;

[56] Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2000), "The Generalized Dynamnic Factor Model: Identification and Estimation," The Review of Economics and Statistics, 82, 540-552;

[57] Forni, M., Hallin, M., Lippi, M., Reichlin, L. (2001) "Coincident and leading indicators for the Euro area", The Economic Journal, 111(471), 62-85;

[58] Friedman, M. (1962) "The interpolation of time series by related series", Journal of the American Statistical Association 57.300 : 729-757;

[59] Geman, H., Kharoubi, C. (2008) "WTI crude oil Futures in portfolio diversification: the time-to-maturity effect", Journal of Banking and Finance 32, 2553–2559;

[60] Geweke, J. (1977) "The Dynamic Factor Analysis of Economic Time Series", in Latent Variables in Socio-Economic Models, eds. D. J. Aigner and A. S. Goldberger, Amsterdam: North-Holland, Ch. 19;

[61] Giannone, D., Reichlin, L., Sala, L. (2005a) "Monetary policy in real time", Macroeconomic Annual 19, 161–200;

[62] Giannone, D., Reichlin, L., Sala, L. (2005b) "VARs, common factors and the empirical validation of equilibrium business cycle models", Journal of Econometrics 127(1), 257–279;

[63] Golinelli, R., Parigi, G. (2007) "The use of monthly indicators to forecast quarterly GDP in the short run: an application to the G7 countries", Journal of Forecasting 26.2 : 77-94;

[64] Gordon, R. H. (1988) "Discussion of taxation and international competitiveness", by Lawrence H Summers, in: Jacob A. Frenkel, ed., International aspects of fiscal policy (University of Chicago Press, Chicago);

[65] Gorton, G. B., Geert-Rouwenhorst, K. (2006) "Facts and Fantasies about Commodity Futures", Financial Analysts Journal 62(2): 47–68;

[66] Hamilton, J. D. (2009a) "Understanding Crude Oil Prices", Energy Journal 20(2):179–206.

[67] Hansen, P. R., Lunde, A., Nason, J. M. (2003) "Choosing the Best Volatility Models: The Model Confidence Set Approach", Oxford Bulletin of Economics and Statistics, 65, 839–861;

[68] Hansen, P. R., Lunde, A., Nason, J. M. (2011) "The Model Con dence Set", Econometrica, 79, 453 497;

[69] Hansen, P. R. (2005) "A test for superio predictive ability", Journal of Business and Economic Statistics, 23, 365-380;

[70] Hazuka, T. B. (1984) "Consumption betas and backwardation in commodity markets", Journal of Finance 39 (July): 647-55;

[71] Ipatova, E., Trapani, L. (2012) "Large Dimentional Panel Interpolation Using EM algorithm", mimio;

[72] Kapetanios, G., Marcellino, M. (2006a) "Factor-GMM estimation with large sets of possibly weak instruments", Unpublished Manuscript;

[73] Kat , H. M., Oomen, R. C. A. (2006) "What Every Investor Should Know About Commodities. Part II: Multivariate Returns Analysis", Alternative investment research centre, working paper #0033;

[74] Kaufmann, K. (2011) "The role of market fundamentals and speculation in recentprice changes for crude oil", Energy Policy 39, 105-115;

[75] Kilian, L. (2008a) "Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market", unpublished manuscript, University of Michigan;

[76] Kilian, L. (2008b) "The Economic Effects of Energy Price Shocks", unpublished manuscript, University of Michigan;

[77] Kilian, L., Dan, M. (2010) "The Role of Inventories and Speculative Trading in the Global Market for Crude Oil", Working Paper, Department of Economics, University of Michigan;

[78] King, G. (2013) " Advanced quantitative research methodology. Harward Lecture Notes". Available at: http://projects.iq.harvard.edu/files/gov2001/files/eviltlk.pdf

[79] Kitchen, J., Monaco,R. (2003) "Real-time forecasting in practice: the US treasury staff's real-time GDP forecast system", Business Economics 38.4 : 10-28;

[80] Krugman, P. (2008) "More on Oil and Speculation", New York Times, May 13. http://krugman.blogs.nytimes.com/2008/05/13/more-on-oil-and-speculation;

[81] Lawley, D. N., Maxwell, A. E. (1962) "Factor analysis as a statistical method", The Statistician, 209-229;

[82] Lawley, D. N., Maxwell, A. E. (1971) "Factor Analysis in a Statistical Method", London: Butterworth;

[83] Malone, T. W., Yates, J., Benjamin, R. I. (1987b) "Electronic markets and electronic hierarchies Commun", ACM 30, 484-497;

[84] Marcellino, M., Schumacher, C., Deutsche Bundesbank (2007) "Factor nowcasting of German GDP with ragged-edge data: A model comparison using MIDAS projections", Technical report, Bundesbank Discussion Paper, Series 1, 34;

[85] Marcellino, M., Stock, J, H., Watson, M. (2003) "Macroeconomic forecasting in the Euro area: Country specific versus Euro wide information", European Economic Review 47, 1–18;

[86] Mariano, R. S., Murasawa, Y. (2002) "A new coincident index of business cycles based on monthly and quarterly series", Journal of Applied Econometrics, 18(4), 427-443;

[87] Mariano, R. S., Murasawa, Y. (2010) "A Coincident Index, Common Factors, and Monthly Real GDP", Oxford Bulletin of Economics and Statistics, 72(1), 27-46;

[88] Masters, M. W. (2008) "Testimony before the Committee on Homeland Security and Government Affairs", U.S. Senate. May 20. http://hsgac.senate.gov/public/_files/052008Masters.pdf;

[89] Mitchell, J. (2005) "An Indicator of Monthly GDP and an Early Estimate of Quarterly GDP Growth", The Economic Journal 115.501: F108-F129;

[90] Mittnik, S., Zadrozny, P. (2005) "Forecasting quarterly German GDP at monthly intervals using monthly Ifo business conditions data", Ifo survey data in business cycle and monetary policy analysis: 19-48;

[91] Mjelde, J. W., Bessler, D. A. (2009) "Market integration among electricity markets and their major fuel source markets", Energy Economics 31 (3), 482-491;

[92] Mou, Y. (2010) "Limits to Arbitrage and Commodity Index Investment: Front-Running the Goldman Roll", working paper, Columbia University;

[93] Nielsen, M., Morin, L. (2011) "FCVAR model.m: A Matlab software package for estimation and testing in the fractionally cointegrated VAR model", QED working paper 1273, Queen's University;

[94] Nunes, L., C. (2005) "Nowcasting quarterly GDP growth in a monthly coincident indicator model", Journal of Forecasting 24.8: 575-592;

[95] O'Dea, D. (1975) "Cyclical Indicators for the postwar British economy", Cambridge: Cambridge University Press. NIESR Occasional Paper # 28;

[96] Onatski, A. (2009) "Testing hypotheses about the number of factors in large factor models", Econometrica, 77, 1447-1479;

[97] Onatski, A. (2010) "Determining the number of factors from empirical distribution of eigenvalues", The Review of Economics and Statistics, 92(4), 1004-1016;

[98] Parigi, G., Schlitzer, G. (1995) "Quarterly forecasts of the Italian business cycle by means of monthly economic indicators", Journal of Forecasting 14.2: 117-141;

[99] Perron, P., Ng, S. (2001) "Lag Length Selection and the Construction of Unit Root Tests with Good Size and Power", Econometrica Vol. 69, No. 6, pp. 1519-1554;

[100] Phillips, P. C. B., Perron, P. (1988) "Testing for a Unit Root in Time Series Regression", Biometrika , Vol. 75, No. 2, pp. 335-346;

[101] Pirrong, C. (2008) "Restricting Speculation Will not Reduce Oil Prices", The Wall Street Journal. Available: http://online.wsj.com/article/SB121573804530744613.html;

[102] Poon S. H., Granger, C. (2003) "Forecasting volatility in financial markets: a review", Journal of Economic Literature 41: 478–539;

[103] Power, G. J., Turvey, C. G. (2011) "Revealing the impact of index traders on commodity futures markets", Applied Economics Letters, 2011, 18, 621–626;

[104] Rencher, A. (2002) "Methods of Multivariate Analysis", Second Edition published by John Wiley & Sons, Inc;

[105] Roodman, D. (2009b) "A Short Note on the Theme of too Many Instruments", Oxford Bulletin of Economics and Statistics, 71(1),135-158;

[106] Ross, S. A. (1976) "The arbitrage theory of capital asset pricing", Rodney L. White Center for Financial Research, University of Pennsylvania, The Wharton School;

[107] Sanders, A., Dwight R., Irwin, S. H. (2008) "Futures Imperfect", New York Times, http://www.nytimes.com/2008/07/20/opinion/20irwinsanders .html?_r=1&oref=slogin;

[108] Sargent, T. J., Sims, C. A. (1977) "Business Cycle Modeling Without Pretending to Have Too Much A Priori Economic Theory", in New Methods in Business Cycle Research, eds. C. Sims et al., Minneapolis: Federal Reserve Bank of Minneapolis;

[109] Schumacher, C., Breitung, J. (2008) "Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data", International Journal of Forecasting 24.3: 386-398;

[110] Schumacher, C. (2005) "Forecasting german GDP using alternative factor models based on large datasets", Bundesbank Discussion Paper 24-2005;

[111] Sharpe, W. F. (1964) "Capital assets prices: A theory of market equilibrium under conditions of risk", Journal of Finance, Vol. 19, pp. 425-442;

[112] Smith, J. L. (2009) "World Oil: Market or Mayhem?", Journal of Economic Perspectives 23(3): 145–64;

[113] Stock, J. H., Watson, M. W. (2002b) "Macroeconomic forecasting using diffusion indexes", Journal of Business and Economic Statistics 20(2), 147–162;

[114] Stock, J. H., Watson, M. W. (1998) "Diffusion indexes", (No. w6702) National Bureau of Economic Research;

[115] Stock, J. H., Watson, M. W. (2002a) "Forecasting using principal components from a large number of predictors", Journal of the American statistical association 97.460: 1167-1179;

[116] Stock, J. H., Watson, M. W. (2005) "Implications of dynamic factor models for VAR analysis", NBER Working Paper 11467;

[117] Stock, J. H., Watson, M. W. (2006) "Forecasting with many predictors", Handbook of economic forecasting: 515-554;

[118] Stock, J. H., Watson, M. W. (1989) "New Indexes of Coincident and Leading Economic Indicators", NBER Macroeconomics Annual, 351-393;

[119] Stock, J. H., Watson, M. W. (1999) "Forecasting inflation", Journal of Monetary Economics, 44(2), 293-335;

[120] Stone, R. (1947) "On the interdependence of blocks of transactions", Supplement to the Journal of the Royal Statistical Society 9.1: 1-45;

[121] Trapani, L. (2012a) "On bootstrapping panel factor series", Journal of Econometrics, forthcoming;

[122] Trapani, L. (2012b) "On bootstrapping panel factor series - Extended version", Available at SSRN: http://ssrn.com/abstract=2062183;

[123] Trehan, B. (1989) "Forecasting Growth in Current Quarter Real GNP", Federal Reserve Bank of S. Francisco Economic Review, Winter, pp. 39-52;

[124] Wallis, K. F. (1986) "Forecasting with an econometric model: The 'ragged edge'problem", Journal of Forecasting 5.1: 1-13;

[125] White, H. (2000) "A Reality Check for Data Snooping", Econometrica, 68, 1097-1126;

[126] Working, H. (1949) "The Theory of the Price of Storage", American Economic Review 39: 1254-62.

[127] Zagaglia, P. (2010) "Macroeconomic factors and oil futures prices: A data-rich model", Energy Economics 32 409-417;

[128] Zhang, Y. J., Fan, Y., Tsai, H. T., Wei, Y. M. (2008) "Spillover effect of US dollar exchange rate on oil prices", Journal of PolicyModeling, 30,973–991;

Last Page