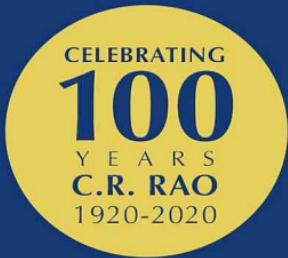


handbook of statistics 42

Financial, Macro and Micro
Econometrics Using R

Edited by
Hrishikesh D. Vinod
C.R. Rao



Handbook of Statistics

Series Editor

C.R. Rao

C.R. Rao AIMSCS, University of Hyderabad Campus,
Hyderabad, India

Arni S.R. Srinivasa Rao

Medical College of Georgia, Augusta University, United States

North-Holland is an imprint of Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

Copyright © 2020 Elsevier B.V. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-12-820250-0

ISSN: 0169-7161

For information on all North-Holland publications
visit our website at <https://www.elsevier.com/books-and-journals>

Publisher: Zoe Kruze
Acquisition Editor: Sam Mahfoudh
Editorial Project Manager: Peter Llewellyn
Production Project Manager: Vignesh Tamil
Cover Designer: Alan Studholme

Typeset by SPi Global, India



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Contributors

Numbers in Parentheses indicate the pages on which the author's contributions begin.

Lekha S. Chakraborty (155), National Institute of Public Finance and Policy,
New Delhi, India; Levy Economics Institute of Bard College, New York, NY,
United States

Jianghao Chu (81), Department of Economics, University of California, Riverside,
CA, United States

Giancarlo Ferrara (299), Department of Economics, Business and Statistics,
University of Palermo, Palermo; SOSE—Soluzioni per il Sistema Economico SpA,
Rome, Italy

Eric Ghysels (117), Department of Economics, UNC Chapel Hill; Department of
Finance, Kenan-Flagler Business School and CEPR Fellow, Chapel Hill, NC,
United States

Honey Karun (155), Senior Resident Representative Office, International Monetary
Fund, New Delhi, India

Virmantas Kvedaras (117), Joint Research Centre, European Commission, Brussels,
Belgium

Tae-Hwy Lee (81), Department of Economics, University of California, Riverside,
CA, United States

Roberto S. Mariano (185), Department of Economics, University of Pennsylvania,
Philadelphia, PA, United States

Arpita Mukherjee (3), Department of Economics, Rutgers University,
New Brunswick, NJ, United States

Suleyman Ozmucur (185), Department of Economics, University of Pennsylvania,
Philadelphia, PA, United States

Weijia Peng (3), Department of Economics, Rutgers University, New Brunswick, NJ,
United States

Peter C.B. Phillips (61), Department of Economics, Yale University, New Haven, CT,
United States; Department of Economics, University of Auckland, Auckland, New
Zealand

Shuping Shi (61), Department of Economics, Macquarie University, North Ryde,
NSW, Australia

Robin C. Sickles (267), Department of Economics, Rice University, Houston, TX,
United States

Wonho Song (267), School of Economics, Chung-Ang University, Seoul, Republic of Korea

Matthieu Stigler (229), Department of Agricultural and Resource Economics, University of California, Davis, CA, United States

Norman R. Swanson (3), Department of Economics, Rutgers University, New Brunswick, NJ, United States

Aman Ullah (81), Department of Economics, University of California, Riverside, CA, United States

Hrishikesh Vinod (155), Fordham University, New York, NY, United States

Xiye Yang (3), Department of Economics, Rutgers University, New Brunswick, NJ, United States

Valentin Zelenyuk (267), School of Economics, University of Queensland, Brisbane, QLD, Australia

Vaidotas Zemlys-Balevičius (117), Institute of Applied Mathematics, Vilnius University, Vilnius, Lithuania

Preface

As with earlier volumes in this series, volume 42 of *Handbook of Statistics* with the subtitle “Financial, Macro and Micro Econometrics Using R” provides state-of-the-art information on important topics in Econometrics, a branch of Economics concerned with quantitative methods. This handbook, a companion of volume 41 with the subtitle “Conceptual Econometrics Using R,” also covers a great many conceptual topics of practical interest to quantitative scientists, especially in Economics and Finance.

The book has uniquely broad coverage with all chapter authors providing practical R software tools for implementing their research results. Despite some overlap, we divide the chapters into three parts. We list the three parts while retaining the nine chapter numbers as:

1. *Finance*

- (1) Arpita Mukherjee, Weijia Peng, Norman R. Swanson and Xiye Yang survey big data models in finance with emphasis on volatility.
- (2) Peter C. B. Phillips and Shuping Shi provide new tools for real-time monitoring of asset market bubbles and crises, important for regulators and investors. They also develop a new bootstrap for statistical inference.
- (3) Jianghao Chu, Tae-Hwy Lee and Aman Ullah work with modern extensions of discrete AdaBoost algorithms adapted for machine learning, classification, and turning point predictions.

2. *Macro Econometrics*

- (4) Eric Ghysels, Virmantas Kvedaras, and Vaidotas Zemlys-Balevičius describe general tools for single equation regressions when variables have mixed set of frequencies (e.g., daily, weekly, yearly).
- (5) Hrishikesh Vinod, Honey Karun, and Lekha S. Chakraborty use novel statistical tools suitable for short time series showing that government investment in India does not “crowd out” private investments.
- (6) Roberto S. Mariano and Suleyman Ozmucur survey data-parsimonious mixed-frequency dynamic latent factor model (MF-DLFM), current quarterly model (CQM), and mixed data sampling (MIDAS) regressions, illustrated by forecasts of Philippine GDP and Inflation. The motivation is to exploit “breaking news” data from relevant high-frequency indicators in updating low frequency (quarterly) forecasts.

- (7) Matthieu Stigler reviews newer research on threshold cointegration allowing for asymmetric and nonlinear adjustment among nonstationary variables. This allows delayed correction of deviations from long-run economic equilibria—only after they exceed a threshold—using R package “*tsDyn*.”

3. *Micro Econometrics*

- (8) Robin C. Sickles, Wonho Song, and Valentin Zelenyuk discuss R tools for “Stochastic Frontier Analysis,” “Data Envelopment Analysis,” and “Free Disposable Hull.”
- (9) Giancarlo Ferrara focuses on stochastic frontier model extensions involving generalized additive models, contextual variables, and spatial external factors.

All chapters are authored by distinguished researchers. Most senior authors have received professional honors, such as being elected “Fellows” of the *Journal of Econometrics* or of the *Econometric Society*.

The intended audience is not only students, teachers, and researchers in various industries and sciences but also profit and nonprofit business decision makers and government policymakers. The wide variety of applications of statistical methodology should be of interest to researchers in all quantitative fields in both natural and social sciences and engineering.

A unique feature of this volume is that all included chapters provide not only a review of the newer theory but describe ways of implementing authors’ new ideas using free R software. Also, the writing style is user-friendly and includes descriptions and links to resources for practical implementations on a *free* open source R, allowing readers to not only use the tools on their own data but also providing a jump start for understanding the state of the art. Open source allows reproducible research and opportunity for anyone to extend the toolbox.

According to a usage dating back to Victorian England, the phrase “The three R’s” describes basic skills taught in schools: Reading, wRiting, and aRithmetic. In the 21st century, we should add R software as the *fourth R*, which is fast becoming an equally basic skill. Unfortunately, some economists are continuing to rely on expensive copyrighted commercial software which not only needs expensive updating but also hides many internal computational algorithms from critical public evaluation for robustness, speed, and accuracy. Users of open source software routinely work with the latest updated versions. This saves time, resources, and effort needed in deciding whether the improvements in the latest update are worth the price and arranging to pay for it.

In teaching undergraduate statistics classes one of us (Vinod) introduces students to R as a convenient calculator, where they can name numerical vector or matrix objects for easy manipulation by name. Starting with the convenience of not having to use Normal or Binomial tables, students begin to appreciate and enjoy the enormous power of R for learning and analyzing quantitative data.

There are over 16,045 free R packages, contributed and maintained by researchers from around the world, which can be searched at <https://mran.microsoft.com/packages>. In short, R has a huge and powerful ecosystem. Students soon learn that if a statistical technique exists, there is most likely an R package which has already implemented it. The plotting functions in R are excellent and easy to use, with the ability to create animations, interactive charts and superimpose statistical information on geographical maps, including the ability to indicate dynamically changing facts. R is able to work with other programming languages including Fortran, Java, C++, and others. R is accessible in the sense that one does not need to have formal training in computer science to write R programs for general use.

For reviewing the papers we thank: Peter R. Hansen (University of North Carolina at Chapel Hill), Shujie Ma (University of California, Riverside), Aaron Smith (University of California, Davis), Tayyeb Shabbir (California State University Dominguez Hills, Carson, CA), Andreas Bauer (IMF Senior Resident Representative, New Delhi, India), José Dias Curto (ISCTE - Instituto Universitario de Lisboa, Portugal), Ruey S. Tsay (Booth School of Business, University of Chicago), Alessandro Magrini (University of Florence, Italy), Jae H. Kim (La Trobe University, Australia), In Choi (Sogang University, Korea), among others.

A common thread in all chapters in this handbook is that all authors of this volume have taken extra effort to make their research implementable in R. We are grateful to our authors as well as many anonymous researchers who have refereed the papers and made valuable suggestions to improve the chapters. We also thank Peter Llewellyn, Kari Naveen, Vignesh Tamilselvvanagnesh, Arni S.R. Srinivasa Rao, Sam Mahfoudh, Alina Cleju, and others connected with Elsevier's editorial offices.

**Hrishikesh D. Vinod
C.R. Rao**

Chapter 1

Financial econometrics and big data: A survey of volatility estimators and tests for the presence of jumps and co-jumps

Arpita Mukherjee*, Weijia Peng, Norman R. Swanson and Xiye Yang

Department of Economics, Rutgers University, New Brunswick, NJ, United States

*Corresponding author: e-mail: am1832@economics.rutgers.edu

Abstract

In recent years, the field of financial econometrics has seen tremendous gains in the amount of data available for use in modeling and prediction. Much of this data is very high frequency, and even “tick-based,” and hence falls into the category of what might be termed “big data.” The availability of such data, particularly that available at high frequency on an intra-day basis, has spurred numerous theoretical advances in the areas of volatility/risk estimation and modeling. In this chapter, we discuss key such advances, beginning with a survey of numerous nonparametric estimators of integrated volatility. Thereafter, we discuss testing for jumps using said estimators. Finally, we discuss recent advances in testing for co-jumps. Such co-jumps are important for a number of reasons. For example, the presence of co-jumps, in contexts where data has been partitioned into continuous and discontinuous (jump) components, is indicative of (near) instantaneous transmission of financial shocks across different sectors and companies in the markets; and hence represents a type of systemic risk. Additionally, the presence of co-jumps across sectors, say, suggests that if jumps can be predicted in one sector, then such predictions may have useful information for

 This chapter has been prepared for inclusion in the Handbook of Statistics. The authors are grateful to the editors, and in particular to H.D. Vinod, who has provided invaluable assistance during the preparation of this chapter. Thanks are also owed to Yacine Aït-Sahalia, Valentina Corradi, Frank Diebold, Rob Engle, Jianqing Fan, Eric Ghysels, Yuan Liao, Bruce Mizrach, and Markus Pelger for discussions that have aided in much of the authors’ own research that is discussed herein.

modeling variables such as returns and volatility in another sector. As an illustration of the methods discussed in this chapter, we carry out an empirical analysis of DOW and NASDAQ stock price returns.

Keywords: Financial econometrics, Integrated volatility, Nonparametric estimator, Continuous time model, Jumps, Co-jumps, Big data, High-frequency data

JEL classification: C22, C52, C53, C58

1 Introduction

The importance of integrated volatility, jumps, and co-jumps in the financial econometrics literature and in terms of successful risk management by investors is quite obvious now, given the amount of research that has gone into this field. Measures of integrated volatility are crucial given the advent of numerous volatility-based derivative products traded in financial markets while tests for jumps are essential in modeling and predicting volatility and returns. Tests of co-jumps on the other hand are meaningful indicators of transmission of financial shocks across different sectors, companies, and markets. The rationale behind this chapter is to discuss some of recent advances in jump and co-jump testing methodology and measurement of integrated volatility, and the properties thereof, in a way which would help both researchers and practitioners in application of such econometric methods in finance. We begin by surveying the most widely used integrated volatility measures, jump and co-jump tests, followed by an empirical analysis using high-frequency intra-day stock prices of DOW 30 companies and ETFs.

Daily integrated volatility is unobservable. Econometricians have developed numerous measures which estimate price fluctuations in a variety of ways. One of the earliest measures is the *realized volatility (RV)* in [Andersen et al. \(2001\)](#). However, this measure does not separate jump variation from variation due to continuous components. [Barndorff-Nielsen and Shephard \(2004\)](#) use the product of adjacent intra-day returns to develop jump robust measures *bipower variation (BPV)* and *tripower variation (TPV)*. One of the more recent techniques of separating out the jump component is the truncation methodology which essentially eliminates returns which are above a given threshold as in [Corsi et al. \(2010\)](#) and [Ait-Sahalia et al. \(2009\)](#). One important caveat of high-frequency data is the existence of market microstructure noise which creates a bias in the estimation procedure. [Zhang et al. \(2005, 2006\)](#) and [Kalnina and Linton \(2008\)](#) solved this problem with noise robust volatility estimators.

In [Duong and Swanson \(2011\)](#), the authors find that 22.8% of the days during the 1993–2000 period had jumps while 9.4% of the days during the 2001–2008 period had jumps. The existence of jumps in financial markets is obvious, which has led many researches to develop techniques which can test for jumps. Jump diffusion is pivotal in analyzing asset movement in financial econometrics and developing jump tests to identify jumps has been

the focus for many theoretical econometricians in past few years. Using the ratio of BPV and estimated quadratic variation, [Barndorff-Nielsen and Shephard \(2006\)](#) construct a nonparametric test for the existence of jumps. [Lee and Mykland \(2007\)](#) on the other hand propose tests to detect the exact timing of jumps at the intra-day level while [Jiang and Oomen \(2008\)](#) provide a “swap variance” approach to detect the presence of jumps. Instead of the more widely used “fixed time span” tests, [Corradi et al. \(2014, 2018\)](#) develop “long time span” jump test, building on earlier work by [Aït-Sahalia \(2002\)](#).

Co-jump tests which are instrumental in identifying systemic risk across multiple sectors and markets are relatively new in the literature. Co-jumps reflect market correlation and have important implication for portfolio management and risk hedging. There are tests which utilize univariate jump tests to identify co-jumps among multivariate processes ([Gilder et al., 2014](#)), while co-jump tests can also be directly applied to multiple price processes (see, e.g., [Jacod and Todorov, 2009](#), [Bandi and Reno, 2016](#), [Bibinger and Winkelmann, 2015](#), [Caporin et al., 2017](#)). [Gnabo et al. \(2014\)](#) propose a co-jump test based on bootstrapping methods, [Bandi and Reno \(2016\)](#) develop a nonparametric infinitesimal moments method to detect co-jumps between asset returns and volatilities and [Caporin et al. \(2017\)](#) build a co-jump test based on the comparison between smoothed realized variance and smoothed random realized variation.

As an illustration of the aforementioned testing methodologies and estimation procedures, an empirical analysis is carried out using high-frequency intra-day stock prices of six DOW 30 companies and ETFs which include The Boeing Company (BA), Exxon Mobile Corporation (XOM), Johnson & Johnson (JNJ), JPMorgan Chase & Co. (JPM), Microsoft Corporation (MSFT), and Walmart Inc. (WMT) and two SPDR sector ETFs XLE & XLK. We use three jump tests; Aït-Sahalia and Jacod (ASJ) test ([Aït-Sahalia et al., 2009](#)), BNS test ([Barndorff-Nielsen and Shephard, 2006](#)), and Lee and Mykland (LM) test ([Lee and Mykland, 2007](#)). In terms of co-jump tests we use JT test ([Jacod and Todorov, 2009](#)), BLT test ([Bollerslev et al., 2008](#)), and GST co-exceedance rule ([Gilder et al., 2014](#)). For estimation of integrated volatility we make use of RV ([Andersen et al., 2001](#)), BPV and TPV ([Barndorff-Nielsen and Shephard, 2004](#)), *truncated realized volatility* (TRV) ([Aït-Sahalia et al., 2009](#)), and *MedRV* and *MinRV* ([Andersen et al., 2012](#)). In our findings, we report the volatility movement of the different stocks and ETFs, percentage of days identified as having jumps and co-jumps, kernel density plots of the different jump and co-jump test statistics as well the proportion of jump variation to the total variation in the asset prices.

The important empirical findings can be summarized as follows. Over the entire sample period JPMorgan has the highest and Johnson & Johnson has the lowest mean estimated integrated volatility. Among all the volatility measures, BPV reports the lowest mean volatility estimate while RV reports the highest mean volatility estimate for any given stock or ETF. This can be explained by the fact that in the presence of frequent jumps, RV overestimates

integrated volatility. All individual stocks achieve their highest volatility in the fourth quarter of 2008 during the financial crisis. XLK sector ETF has the largest percentage of jump days (38%) and ratio of jump to total variation (45%) among all other ETFs and individual stocks. BNS jump test detected more jumps and reported a larger percentage of jump days when compared with the other two jump tests. When the sampling frequency is reduced from 1 to 5 min, the ASJ jump test reports lesser number of jumps as well as smaller proportion of jump to total variation in the sample data. We detect co-jumps between Exxon & JPMorgan, Exxon & Microsoft, Exxon & XLE, JPMorgan & Microsoft, Microsoft & XLK, and XLE & XLK through JT co-jump test and the GST co-exceedance rule. The results show that the percentage of co-jump days range from 0.4% to 2.5% for JT co-jump test and from 2.8% to 9.5% for the GST co-exceedance rule. The higher percentage of co-jump days in case of the co-exceedance rule, which uses the results at the intersection of BNS and LM jump tests, could be because the test has a large false rejection rate. We use BLT co-jump test to detect co-jumps among six stocks including Boeing, Exxon, Johnson & Johnson, JPMorgan, Microsoft, and Walmart. The percentage of co-jumps days is 0.2% during financial crisis period and 0.1% after financial crisis period.

The rest of the chapter is organized as follows. [Section 2](#) gives the theoretical background and setup. [Sections 3–5](#) give detailed descriptions of the different integrated volatility measures, jump tests, and co-jump tests, respectively. [Section 6](#) discusses the empirical methodology and reports the findings. Finally [Section 7](#) concludes this chapter. Accompanying R code for the measures and tests is provided in [Appendix](#).

2 Setup

We represent the log-price of a financial asset at continuous time t , as Y_t . It is assumed that the log-price is a Brownian semimartingale process with jumps and it can be denoted as^a:

$$Y_t = Y_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s + J_t \quad (1)$$

In (1) μ_s the drift term is a predictable process, σ_s the diffusion term is a càdlàg process, W_s is a standard Brownian motion, and J_t is a pure jump process. J_t can be defined as the sum of all discontinuous log-price movements up to time t ,

$$J_t = \sum_{s \leq t} \Delta Y_s \quad (2)$$

^aWe follow the setup and notation as in [Corradi et al. \(2011\)](#) and [Mukherjee and Swanson \(2018\)](#).

When this jump component is a finite activity jump process, i.e., a compound Poisson process (CPP), then

$$J_t = \sum_{j=1}^{N_t} \xi_j \quad (3)$$

where N_t is a Poisson process with intensity λ , the jumps occur at the corresponding times given as $(\tau_j)_{j=1,\dots,N_t}$, and ξ_j refers to *i.i.d* random variables measuring the size of jumps at time τ_j . The finite activity jump assumption has been widely used in financial econometrics literature. Log-price Y_t can be decomposed into a continuous Brownian component Y_t^c and a discontinuous component Y_t^d (due to jumps). The “true variance” of process Y_t can be given as:

$$QV_t = [Y, Y]_t = [Y, Y]_t^c + [Y, Y]_t^d \quad (4)$$

where QV stands for quadratic variation. The variation due to the continuous component is

$$[Y, Y]_t^c = \int_0^t \sigma_s^2 ds, \quad (5)$$

and the variation due to the discontinuous jump component is

$$[Y, Y]_t^d = \sum_{j=1}^{N_t} \xi_j^2 \quad (6)$$

Integrated volatility which is the continuous part of QV is denoted as:

$$IV_t = \int_{t-1}^t \sigma_s^2 ds, \quad t = 1, \dots, T \quad (7)$$

where IV is the (daily) integrated volatility at day t . Since IV is unobservable, different realized measures of integrated volatility are used as its substitute. The presence of market frictions in high-frequency financial data has been documented in recent literature. To take care of this, the observed log-price process X can then be given as

$$X = Y + \epsilon \quad (8)$$

where Y is the latent log price and ϵ captures market microstructure noise. We consider M equi-spaced intra-daily observations for each of T days for process X which leads to a total of MT observations, i.e.,

$$X_{t+j/M} = Y_{t+j/M} + \epsilon_{t+j/M}, \quad t = 0, \dots, T \text{ and } j = 1, \dots, M \quad (9)$$

where ϵ follows a zero mean independent process. The intra-daily return or increment of process X follows:

$$\Delta_j X = X_{t+(j+1)/M} - X_{t+j/M} \quad (10)$$

The noise containing realized measure, RM of the integrated volatility is computed using process X given in (9) and can be expressed as the sum of IV and measurement error N , i.e.,

$$RM_{t,M} = IV_t + N_{t,M} \quad (11)$$

RM can be used to estimate IV if k th moment of the measurement error decays to zero at a fast enough rate or there exists a sequence b_M with $b_M \rightarrow \infty$ such that $E(|N_{t,M}|^k) = O(b_M^{-K/2})$, for some $k \geq 2$.

3 Realized measures of integrated volatility

Volatility measures variation in the asset prices and thus can be regarded as an indicator of risk. Accurate volatility estimation is very important in both asset allocation and risk management. Since volatility is inherently unobservable, the first two types of parametric models developed to estimate the latent volatility were continuous time (e.g., stochastic volatility) and discrete time models (e.g., ARCH-GARCH models). However, these parametric models have been proven to be misspecified in capturing volatilities implied by option pricing and other financial return variables. With the availability of high-frequency data, a series of nonparametric models have been proposed to examine integrated volatility at intra-day level. [Andersen et al. \(2001\)](#) first introduce a nonparametric volatility measure, termed *realized volatility* by summing over intra-day squared returns. The authors showed that RV is an error free estimator of integrated volatility in the absence of noise and jumps. When the sampling frequency of the data is relatively high, microstructure noise creates a bias in the volatility estimation procedure. [Zhang et al. \(2005, 2006\)](#) and [Kalnina and Linton \(2008\)](#) solve this problem with microstructure noise robust estimators based on subsampling with multiple time scales. [Barndorff-Nielsen et al. \(2008, 2011\)](#), on the other hand, use kernel-based estimators to account for the microstructure noise in finely sampled data. When estimating integrated volatility in the presence of jumps within the underlying price process, jump components should be separated from the quadratic variation. [Barndorff-Nielsen et al. \(2003\)](#) and [Barndorff-Nielsen and Shephard \(2004\)](#) provide asymptotically unbiased integrated volatility estimators, the *BPVs* and *TPVs*, which are robust to the presence of jumps. [Aït-Sahalia et al. \(2009\)](#) propose a threshold method to identify and truncate jumps and further develop a consistent nonparametric jump robust estimator of the integrated volatility. [Corsi et al. \(2010\)](#) introduce threshold bipower variation (*TBPV*) by combining the concepts from [Barndorff-Nielsen et al. \(2003\)](#) and [Mancini \(2009\)](#). [Jacod et al. \(2014\)](#) estimate local volatility by using the empirical characteristic function of the return and then remove bias due to jump variation. When combining both jumps and microstructure noise in the price process, [Fan and Wang \(2007\)](#) propose a wavelet-based multiscale approach to estimate integrated volatility.

Podolskij et al. (2009) design modulated bipower variation (*MBV*), an estimator that filters the impact of microstructure noise then use *BPV* for volatility estimation. Andersen et al. (2012) use the concept of “nearest neighbor truncation” to establish jump and noise robust volatility estimators. On the other hand Brownlees et al. (2016) create truncated two scaled *RV* by adopting a jump signaling indicator as in Mancini (2009) and noise robust subsampling as in Zhang et al. (2005). In addition to the above mentioned work, discussion regarding nonparametric estimation of integrated volatility and functionals of volatility can also be found in Barndorff-Nielsen et al. (2006), Mykland and Zhang (2009), Todorov and Tauchen (2012), Hautsch and Podolskij (2013), and Jacod et al. (2013, 2017); Jing et al. (2014). What follows in the next section is a detailed review of 12 of the most commonly used integrated volatility measures.^b

3.1 Realized volatility

Realized volatility or *RV* as developed in Andersen et al. (2001) is one of the first empirical measures that used high-frequency intra-day returns to compute daily return variability without having to explicitly model the intra-day data. The authors show that under suitable conditions *RV* is an unbiased and highly efficient estimator of *QV* as in (4). By extension it can be shown that in the absence of jumps or when jumps populate the data infrequently, *RV* converges in probability to *IV* as $M \rightarrow \infty$. It should also be noted that *RV* has been used widely as part of the HAR-RV forecasting models. Here

$$RV_{t,M} = \sum_{j=1}^{M-1} (X_{t+(j+1)/M} - X_{t+j/M})^2 \quad (12)$$

3.2 Realized bipower variation

In Barndorff-Nielsen and Shephard (2004), the authors demonstrate that they could untangle the continuous component of quadratic variation from its discontinuous component (jumps). This led them to develop *BPV*, one of the first asymptotically unbiased estimators of *IV* which was robust to the presence of price jumps. It takes the following form

$$BPV_{t,M} = (\mu_1)^{-2} \sum_{j=2}^{M-1} |\Delta_j X| |\Delta_{j-1} X| \quad (13)$$

where $\Delta_j X$ is the same as in (10) and $\mu_1 = 2^{\frac{1}{2}} \frac{\Gamma(1)}{\Gamma(\frac{1}{2})}$.

^bWe follow the notation and description as in Mukherjee and Swanson (2018).

3.3 Tripower variation

The *BPV* does not allow the consistency of the *IV* estimate to be impacted by finite activity jumps. However, it is subject to finite sample jump distortions or upward bias. To counter this problem, *BPV* is generalized to *TPV* in [Barndorff-Nielsen and Shephard \(2004\)](#), by utilizing products of the (lower order) power of three adjacent intra-day returns. Theoretically speaking, although *TPV* is more efficient, it is also more vulnerable to microstructure noise of the high-frequency return data compared to *BPV*. *TPV* can be given as

$$TPV_{t,M} = \left(\mu_2\right)^{-3} \sum_{j=3}^{M-1} |\Delta_j X|^{2/3} |\Delta_{j-1} X|^{2/3} |\Delta_{j-2} X|^{2/3} \quad (14)$$

where $\Delta_j X$ is the same as in (10) and $\mu_2 = 2^{\frac{1}{3}} \frac{\Gamma(\frac{5}{6})}{\Gamma(\frac{1}{2})}$.

3.4 Two-scale realized volatility

It is found that when the sampling interval of the asset prices is small, microstructure noise issues become more prominent and *RV* ceases to function as a robust volatility estimator. Due to the bias introduced by the market microstructure noise in the finely sampled data, initially longer time horizons are preferred by econometricians. It is found that ignoring microstructure noise works well for intervals more than 10 min. However, sampling over lower frequencies does not quantify and correct the noise effect on volatility estimation. As a solution, *two-scale realized volatility* (*TSRV*) is introduced in [Zhang et al. \(2005\)](#) by combining estimators obtained over two time scales, *avg* and *M*. It forms an unbiased and consistent, microstructure noise robust estimator of *IV* in the absence of jumps. It takes the following form

$$TSRV_{t,M} = [X, X]^{avg} - \frac{1}{K} [X, X]^M \quad (15)$$

where

$$[X, X]^{m_i} = \sum_{j=1}^{m_i-1} (X_{t+((j+1)K+i)/M} - X_{t+(jK+i)/M})^2, \quad i = 1, \dots, K \quad \text{and} \quad m_i = \frac{M}{K} \quad (16)$$

$$[X, X]^{avg} = \frac{1}{K} \sum_{i=1}^K [X, X]^{m_i} \quad (17)$$

$$[X, X]^M = \sum_{j=1}^{M-1} (X_{t+(j+1)/M} - X_{t+j/M})^2 \quad (18)$$

$K = cM^{2/3}$ is the number of subsamples, $\frac{M}{K}$ is subsample size, $c > 0$ is a constant, and M is the number of equi-spaced intra-daily observations.

3.5 Multiscale realized volatility

The *TSRV* estimator is not efficient although it has many desirable properties. The rate of convergence for *TSRV* is not satisfactory, it converges to the true volatility (*IV* in the absence of jumps) only at the rate of $M^{-1/6}$. The *multiscale realized volatility (MSRV)* is proposed in [Zhang et al. \(2006\)](#). This is a microstructure noise robust measure which converged to *IV* (in the absence of jumps) at the rate of $M^{-1/4}$. While *TSRV* uses two time scales, *MSRV* on the other hand uses N different time scales. *MSRV* takes the following form

$$MSRV_{t,M} = \sum_{n=1}^N a_n [X, X]^{(M, K_n)}, \quad n = 1, \dots, N \quad (19)$$

where

$$a_n = 12 \frac{n(n/N - 1/2 - 1/(2N))}{N^2}, \quad \sum_{n=1}^N a_n = 1 \quad \text{and} \quad \sum_{n=1}^N a_n/n = 0 \quad (20)$$

$$[X, X]^{(M, K_n)} = \frac{1}{K_n} \sum_{l=1}^{K_n} \sum_{j=1}^{m_{n,l}-1} (X_{t+((j+1)K_n+l)/M} - X_{t+(jK_n+l)/M})^2 \quad (21)$$

Here $l = 1, \dots, K_n$ and $m_{n,l} = \frac{M}{K_n}$. We take $N = 3, K_1 = 1, K_2 = 2, K_3 = 3$.

3.6 Realized kernel

[Barndorff-Nielsen et al. \(2008\)](#) introduce *realized kernel (RK)* which as the name suggests is a realized kernel type consistent measure of *IV* in the absence of jumps. It is robust to endogenous microstructure noise and for particular choices of weight functions it can be asymptotically equivalent to *TSRV* and *MSRV* estimators, or even more efficient. *RK* can be given as

$$RK_{t,M} = \gamma_0(X) + \sum_{h=1}^H \kappa\left(\frac{h-1}{H}\right) \{\gamma_h(X) + \gamma_{-h}(X)\} \quad (22)$$

where

$$\gamma_h(X) = \sum_{j=1}^{M-1} (X_{t+(j+1)/M} - X_{t+j/M})(X_{t+(j+1-h)/M} - X_{t+(j-h)/M}) \quad (23)$$

Here c is a constant. For our analysis we take a Turkey–Hanning₂ kernel which gives $\kappa(x) = \sin^2\{\pi/2(1-x)^2\}$ and $H = cM^{1/2}$.

3.7 Truncated realized volatility

TRV is one of the first volatility measures that tried to estimate *IV* by identifying when price jumps greater than an adequately defined threshold occurred as in [Aït-Sahalia et al. \(2009\)](#). The truncation level for the jumps are chosen in a data-driven manner; the cutoff level α (given below) is set equal to a particular number times estimated standard deviations of the continuous part of the semimartingale. The price jump robust measure can be given as

$$TRV_{t,M} = \sum_{j=1}^{M-1} |\Delta_j X|^2 1_{\{|\Delta_j X| \leq \alpha \Delta_M^{\sigma}\}} \quad (24)$$

where

$$\alpha = 5 \sqrt{\sum_{j=1}^{M-1} |\Delta_j X|^2 1_{\{|\Delta_j X| \leq \Delta_M^{1/2}\}}} \quad (25)$$

Here $\sigma = 0.47$. $\Delta_M = 1/M$

3.8 Modulated bipower variation

MBV as in [Podolskij et al. \(2009\)](#) consistently estimates *IV* and is robust to both market microstructure noise and finite activity jumps. It takes the following form

$$MBV_{t,M} = \frac{(c_1 c_2 / \mu_1^2) mbv_{t,M} - \vartheta_2 \hat{\omega}^2}{\vartheta_1} \quad (26)$$

where

$$\vartheta_1 = \frac{c_1 (3c_2 - 4 + \max((2 - c_2)^3, 0))}{3(c_2 - 1)^2}, \quad \vartheta_2 = \frac{2 \min((c_2 - 1), 1)}{c_1 (c_2 - 1)^2} \quad (27)$$

$$mbv_{t,M} = \sum_{b=1}^B |\bar{X}_b^{(R)}| |\bar{X}_{b+1}^{(R)}| \quad (28)$$

$$\bar{X}_b^{(R)} = \frac{1}{M/B - R + 1} \sum_{j=(b-1)M/B}^{bM/B - R} (X_{t+(j+R)/M} - X_{t+j/M}) \quad (29)$$

Here $c_1 = 2$, $c_2 = 2.3$, $R \approx c_1 M^{0.5}$, $B = 6$, $\mu_1 = 0.7979$, $\hat{\omega}^2 = \frac{1}{2M} RV_{t,M}$, $RV_{t,M}$ is given by (12).

3.9 Threshold bipower variation

[Corsi et al. \(2010\)](#) introduce a jump robust measure, *TBPV* which is constructed by combining the concepts of *BPV*, and *Threshold Realized Variance* ([Mancini, 2009](#)). The authors show that *TBPV* is robust to the choice of threshold function (v as given below).

$$TBPV_{t,M} = \mu_1^{-2} \sum_{j=2}^{M-1} |\Delta_{j-1}X| |\Delta_j X| I_{\{|\Delta_{j-1}X|^2 \leq v_{j-1}\}} I_{\{|\Delta_j X|^2 \leq v_j\}} \quad (30)$$

where

$$v_j = c_v^2 \hat{V}_j \quad (31)$$

$$\hat{V}_j^z = \frac{\sum_{i=-L}^L \kappa\left(\frac{i}{L}\right) (\Delta_{j+i}X)^2 I_{\{(\Delta_{j+i}X)^2 \leq c_v^2 \hat{V}_{j+i}^{z-1}\}}}{\sum_{i=-L}^L \kappa\left(\frac{i}{L}\right) I_{\{(\Delta_{j+i}X)^2 \leq c_v^2 \hat{V}_{j+i}^{z-1}\}}}. \quad (32)$$

and $\Delta_j X$ is given by (10). Here we take $L = 25, c_v = 3, \hat{V}^0 = +\infty$. v_j is the threshold for removal of large returns at each j . \hat{V}_j^z gives estimated local variance in the presence of jumps at each iteration z for any j . Large returns are removed at each iteration according to $\{(\Delta_j X)^2 \leq c_v^2 \hat{V}_j^{z-1}\}$ and the estimated variance at that iteration is multiplied by c_v^2 to get the threshold for the next iteration. When large returns cannot be removed any more, the iterations stop. Typically z is taken to be 2.

3.10 Subsampled realized kernel

Barndorff-Nielsen et al. (2011) constructed *subsampled realized kernel* (SRK) by combining the concepts of subsampling (Zhang et al., 2005) and RK (Barndorff-Nielsen et al., 2008). The main benefit of subsampling in this context is that it can overpower the inefficiency that stems from the poor selection of kernel weights that might be the case in RK. SRK takes the following form

$$SRK_{t,M} = \frac{1}{S} \sum_{s=1}^S K^s(X) \quad (33)$$

where

$$K^s(X) = \gamma_0^s(X) + \sum_{h=1}^H \kappa\left(\frac{h-1}{H}\right) \{\gamma_h^s(X) + \gamma_{-h}^s(X)\} \quad (34)$$

$$\gamma_h^s(X) = \sum_{j=1}^{M/S} x_j^s x_{j-h}^s \quad (35)$$

$$x_j^s = X_{t+ \left(j + \frac{s-1}{S}\right)/M} - X_{t+ \left(j + \frac{s-1}{S}-1\right)/M} \quad (36)$$

Here the smooth Turkey-Hanning₂ kernel function gives $\kappa(x) = \sin^2\{\pi/2(1-x)^2\}$, $S = 13$ and $H = 3$.

3.11 *MedRV* and *MinRV*

As alternatives to *BPV* and *TPV*, [Andersen et al. \(2012\)](#) provide two alternative measures *MedRV* and *MinRV* which are robust to jumps and/or micro-structure noise by using “nearest neighbor truncation.” The basic concept behind these new measures is that the neighboring returns control the level of truncation of absolute returns. On one hand where *MinRV* compares and takes the minimum of two adjacent absolute returns, *MedRV* takes the median of three adjacent absolute returns and carries out two-sided truncation. Unlike the typical truncated realized measures as in [Corsi et al. \(2010\)](#), these new measures do not have to deal with the selection of an ex ante threshold.

$$\text{MinRV}_{t,M} = \frac{\pi}{\pi-2} \left(\frac{M}{M-1} \right) \sum_{j=1}^{M-1} \min(|\Delta_j X|, |\Delta_{j+1} X|)^2 \quad (37)$$

$$\text{MedRV}_{t,M} = \frac{\pi}{6-4\sqrt{3}+\pi} \left(\frac{M}{M-2} \right) \sum_{j=2}^{M-1} \text{med}(|\Delta_{j-1} X|, |\Delta_j X|, |\Delta_{j+1} X|)^2 \quad (38)$$

where $\Delta_j X$ is given by (10).

4 Jump testing

Jump diffusion has become increasingly important in characterizing dynamic movement of asset prices. Early studies about jump diffusions can be seen in [Andersen et al. \(2002\)](#), [Chernov et al. \(2003\)](#), [Pan \(2002\)](#), and [Eraker et al. \(2003\)](#). Differentiating jumps from continuous process is particularly useful because it has implications for both researchers and practitioners in financial econometrics. Thus, a strand of literature has addressed the methodologies to identify jumps in the discretely sampled financial data. [Ait-Sahalia \(2002\)](#) relies on the transition density to test the existence of jumps under the option pricing model. Focusing on the risk-neutral dynamics of the underlying option prices, [Carr and Wu \(2003\)](#) propose a method to use the convergence rates of option prices to distinguish jumps from continuous process. [Johannes \(2004\)](#) proposes a jump test to identify jump-induced misspecification. However, these tests only use limited low frequency data. With availability of high-frequency data, the mechanism behind jump testing methodology has evolved. [Barndorff-Nielsen and Shephard \(2006\)](#) use the ratio of *BPV* and realized quadratic variation to construct a nonparametric test for the existence of jumps. [Huang and Tauchen \(2005\)](#) design extensive Monte Carlo experiments to evaluate the properties of newly proposed jump tests (see [Andersen et al., 2003](#), [Barndorff-Nielsen and Shephard, 2004](#), [Barndorff-Nielsen and Shephard, 2006](#)). [Lee and Mykland \(2007\)](#) propose tests to detect the exact timing of jumps at the intra-day level while [Jiang and Oomen \(2008\)](#) provide a “swap variance” approach to detect the presence of jumps. [Mancini \(2009\)](#) and [Corsi et al. \(2010\)](#) devise unique

threshold or truncation techniques in their testing methodology. Aït-Sahalia et al. (2009) compare two higher order realized power variations to develop a test statistic for the null hypothesis of no jumps. On the other hand Podolskij and Ziggel (2010) combine the concepts truncated power variation and wild bootstrap to propose a threshold-based jump test. In most of the aforementioned papers, the presence of realized jumps is tested over a “fixed time span.” Corradi et al. (2014, 2018) proposed a “long time span” jump test instead, building on earlier work by Aït-Sahalia (2002). More related work on jump tests, self-excitation, and mutual excitation in realized jumps can be found in Lee et al. (2013), Dungey et al. (2016), and Boswijk et al. (2018). In the next section we discuss six different jump tests which arise from different branches of the jump testing literature.

4.1 Barndorff-Nielsen and Shephard test

To test for the existence of jumps in the sample path of asset prices, Barndorff-Nielsen and Shephard (2006) propose nonparametric Hausman (1978) type tests using the difference between *Realized Quadratic Variation*, an estimator of integrated volatility which is not robust to jumps, and *BPV*, which is a jump robust estimator of integrated volatility. *Realized Quadratic Variation* is considered to be the same as *RV*. The adjusted jump ratio test statistic can be given as:

$$BNS = \frac{M^{1/2}}{\sqrt{9\max\left(1, \frac{QPV}{(\mu_1^2 BPV)^2}\right)}} \left(1 - \frac{BPV}{RV}\right) \xrightarrow{d} N(0, 1) \quad (39)$$

where *BPV* is the same as in (13), *RV* is the same as in (12), $9 = ((\pi^2/4) + \pi - 5) \approx 0.6090$. The realized quadpower variation *QPV* is used to estimate integrated quarticity ($\int_0^t \sigma_s^4 ds$) and can be given as:

$$QPV = M \sum_{j=4}^M |\Delta_j X| |\Delta_{j-1} X| |\Delta_{j-2} X| |\Delta_{j-3} X| \xrightarrow{d} \mu_1^4 \int_0^t \sigma_s^4 ds \quad (40)$$

The authors show that the null hypothesis of no jumps is rejected if the test statistic *Barndorff-Nielsen and Shephard test* (BNS) is significantly positive.

4.2 Lee and Mykland test

Lee and Mykland (2007) use the ratio of realized return to estimated instantaneous volatility, and further construct a nonparametric jump test to detect the exact timing of jumps at the intra-day level. The test statistic which identifies whether there is a jump during $(t + j/M, t + (j + 1)/M]$ can be given as:

$$L_{(t+(j+1)/M)} = \frac{X_{t+(j+1)/M} - \widehat{X}_{t+(j+1)/M}}{\sigma_{t+\widehat{(j+1)}/M}} \quad (41)$$

where

$$\sigma_{t+\widehat{(j+1)}/M}^2 \equiv \frac{1}{K-2} \sum_{i=j-K+1}^{j-2} |X_{t+(i+1)/M} - X_{t+i/M}| |X_{t+i/M} - X_{t+(i-1)/M}| \quad (42)$$

Here K is the window size of a local movement of the process. It is chosen in a way such that the effect of jumps on volatility estimation is eliminated. The authors suggest a value of $K = 10$ when the sampling frequency is 5 min. Thus, it can be asymptotically shown that

$$\frac{\max_{j \in \bar{A}_M} |L_{(t+(j+1)/M)}| - C_M}{S_M} \rightarrow \varepsilon, \text{ as } \Delta t \rightarrow 0, \quad (43)$$

where ε has a cumulative distribution function $P(\varepsilon \leq x) = \exp(-e^{-x})$,

$$C_M = \frac{(2\log M)^{1/2}}{c} - \frac{\log \pi + \log(\log M)}{2c(2\log M)^{1/2}} \text{ and } S_M = \frac{1}{c(2\log M)^{1/2}} \quad (44)$$

M is the number of intra-daily observations, $c \approx 0.7979$ and \bar{A}_M is the set of $j \in \{0, 1, \dots, M\}$ so that there are no jumps in $(t + j/M, t + (j + 1)/M]$.

4.3 Jiang and Oomen test

[Jiang and Oomen \(2008\)](#) compare a jump sensitive variance measure to RV in order to test for jumps. Their idea is based on the fact that in the absence of jumps the accumulated difference between the simple return and log return (called the swap variance) captures one-half of the integrated volatility in the continuous time limit. Consequently it can be stated, in the absence of jumps the difference between swap variance and RV should be zero, while in the presence of jumps the same difference reflects the replication error of variance swap thus detecting jumps. The swap variance can be given as:

$$SV_{t,M} = 2 \sum_{j=1}^{M-1} (\Delta_j P - \Delta_j X) \quad (45)$$

where $Y = \log(P)$ and Y is the same as in (1). $\Delta_j P = \frac{P_{t+(j+1)/M}}{P_{t+j/M}} - 1$ and $\Delta_j X$ is the same as in (10). The three different swap variance tests proposed by the authors can be given as:

(i) The difference test:

$$\frac{M}{\Omega_{SV}} (SV_{t,M} - RV_{t,M}) \xrightarrow{d} N(0, 1) \quad (46)$$

(ii) The logarithmic test:

$$\frac{BPV_{t,M}M}{\Omega_{SV}}(\log(SV_{t,M}) - \log(RV_{t,M})) \xrightarrow{d} N(0, 1) \quad (47)$$

(iii) The ratio test:

$$\frac{BPV_{t,M}M}{\Omega_{SV}}\left(1 - \frac{RV_{t,M}}{SV_{t,M}}\right) \xrightarrow{d} N(0, 1) \quad (48)$$

where $\Omega_{SV} = \frac{\mu_6}{9} \frac{M^3 \mu_{6/p}^{-p}}{M-p+1} \sum_{j=1}^{M-p} \prod_{k=0}^p |\Delta_{j+k} X|^{\frac{6}{p}}$ for $p \in \{1, 2, \dots\}$, $\mu_z = E(|x|^z)$ for $z \sim N(0, 1)$.

4.4 Aït-Sahalia and Jacod test

In [Aït-Sahalia et al. \(2009\)](#), the authors develop a testing methodology for jumps in the (log) price process by comparing two higher order realized power variations with different sampling intervals, $k\Delta$ and Δ , respectively. In this context $\Delta = \frac{1}{M}$, M is the number of intra-daily observations and k is a given integer. The p th order realized power variation can be given as

$$\hat{B}(p, \Delta) = \sum_{j=1}^{M-1} |X_{t+(j+1)/M} - X_{t+j/M}|^p \quad (49)$$

The ratio of the two realized power variations with different sampling intervals takes the following form

$$\hat{S}(p, k, \Delta) = \frac{\hat{B}(p, k\Delta)}{\hat{B}(p, \Delta)} \quad (50)$$

The corresponding jump test statistic can then be defined as:

$$ASJ = \frac{k^{(p/2)-1} - \hat{S}(p, k, \Delta)}{\sqrt{V_{t,M}}} \xrightarrow{d} N(0, 1) \quad (51)$$

where $V_{t,M}$ can be estimated using either a truncation technique as in

$$\hat{V}_{t,M} = \Delta \frac{\hat{A}(2p, \Delta) M(p, k)}{\hat{A}(p, \Delta)^2} \quad (52)$$

where

$$\hat{A}(2p, \Delta) = \frac{\Delta^{1-p/2}}{\mu_p} \sum_{j=1}^{M-1} |X_{t+(j+1)/M} - X_{t+j/M}|^p \mathbf{1}_{\{|X_{t+(j+1)/M} - X_{t+j/M}| \leq \alpha \Delta^\varpi\}} \quad (53)$$

or using multipower variation as in

$$\hat{V}_{t,M} = \Delta \frac{M(p,k) \bar{A}(p/([p]+1), 2[p]+2, \Delta)}{\bar{A}(p/([p]+1), [p]+1, \Delta)^2} \quad (54)$$

where

$$\bar{A}(r,q,\Delta) = \frac{\Delta^{1-q_r/2}}{\mu_r^q} \sum_{j=q}^{M-q+1} \prod_{i=0}^{q-1} |X_{t+(j+i)/M} - X_{t+(j+i-1)/M}|^r, \quad (55)$$

$$M(p,k) = \frac{1}{\mu_p^2} (k^{p-2} (1+k) \mu_{2p} + k^{p-2} (k-1) \mu_p^2 - 2k^{p/2-1} - \mu_{k,p}) \quad (56)$$

and $\mu_r = E(|U|^r)$ and $\mu_{k,p} = E(|U|^p | U + \sqrt{(k-1)V^p}|)$ for $U, V \sim N(0, 1)$. The null hypothesis of no jumps is rejected when the test statistic ASJ is significantly positive.

4.5 Podolskij and Ziggel (PZ) test

In [Podolskij and Ziggel \(2010\)](#) the concept of truncated power variation is used to construct test statistics which diverge to infinity if jumps are present and have a normal distribution otherwise. The jump testing procedure in this chapter is valid (under weak assumptions) for all semimartingales with absolute continuous characteristics and general models for the noise processes. The methodology followed by the authors is a modification of that proposed in [Mancini \(2009\)](#). In particular they consider

$$T(X,p) = M^{\frac{p-1}{2}} \sum_{j=1}^{M-1} |X_{t+(j+1)/M} - X_{t+j/M}|^p (1 - \eta_i 1_{\{|X_{t+(j+1)/M} - X_{t+j/M}| \leq \alpha \Delta^\sigma\}}) \quad (57)$$

where $\{\eta_i\}_{i \in [1, 1/\Delta]}$ is a sequence of positive *i.i.d* random variables. The test statistic has the following form

$$PZ = \frac{T(X,p)}{Var^*(\eta) \hat{A}(2p, \Delta)} \xrightarrow{d} N(0, 1) \quad (58)$$

where $\hat{A}(2p, \Delta)$ is the same as in (53).

4.6 Corradi, Silvapulle, and Swanson test

Building on previous work by [Aït-Sahalia \(2002\)](#) and [Corradi et al. \(2018\)](#) design “long time span” jump tests based on realized third moments or “tricity” for the null hypothesis that the probability of a jump is zero. This jump testing methodology is used to detect jumps by examining the “jump intensity” parameter in the data generating process rather than realized jumps over a “fixed time span.” This test is of immense value when one is interested

in using jump diffusion processes for valuation problems like options pricing and default modeling. Let,

$$\begin{aligned}\hat{\mu}_{3,T,\Delta} = & \frac{1}{T} \sum_{j=1}^{n-1} \left(X_{t+(j+1)/M} - X_{t+j/M} - \frac{X_{t+n/M} - X_{t+1/M}}{n} \right)^3 \\ & - \frac{1}{T^+} \sum_{j=1}^{n^+-1} \left(X_{t+(j+1)/M} - X_{t+j/M} - \frac{X_{t+n^+/M} - X_{t+1/M}}{n^+} \right)^3 \\ & 1\{|X_{t+(j+1)/M} - X_{t+j/M}| \leq \tau(\Delta)\}\end{aligned}\quad (59)$$

where we have n^+ observations over an increasing time span of T^+ , a shrinking discrete sampling interval $\Delta = \frac{1}{M}$, so that $n^+ = \frac{T^+}{\Delta}$, $T^+ \rightarrow \infty$ and $\Delta \rightarrow 0$. $\tau(\Delta)$ is the truncation parameter and one example for the choice of such truncation can be given as follows. If σ_s as in (1) is a square root process, so that all moments exist, we can set $\tau(\Delta) = c\Delta^\eta$ with $\frac{2}{\eta} < \eta < \frac{1}{2}$. The authors define $n = \frac{T}{\Delta} = n^+ - \frac{T^+ - T}{\Delta}$, with $T^+ > T$ and $\frac{T^+}{T} \rightarrow \infty$. Then, the test statistic for the null hypothesis of no jumps can be given as

$$CSS = \frac{T^{1/2}}{\Delta} \hat{\mu}_{3,T,\Delta} \xrightarrow{d} N(0, \omega_0) \quad (60)$$

where ω_0 is defined in Corradi et al. (2018). Since, under the alternative hypothesis of positive jump intensity, the variance of the statistic is of larger order, it is difficult to construct a variance estimator which is consistent under all hypotheses. The authors use a threshold variance estimator, which removes the contribution of the jump component thus developing an estimator for the variance of Corradi, Silvapulle, and Swanson test (CSS) which is consistent under the null hypothesis of no jumps. Thus we have

$$\begin{aligned}\hat{\sigma}_{CSS}^2 = & \frac{1}{\Delta^2} \sum_{j=0}^{n-1} \left(X_{t+(j+1)/M} - X_{t+j/M} - \frac{X_{t+n/M} - X_{t+1/M}}{n} \right)^3 \\ & 1\{|X_{t+(j+1)/M} - X_{t+j/M}| \leq \tau(\Delta)\}\end{aligned}\quad (61)$$

Thus the t -statistic version of the jump test is

$$t_{CSS} = \frac{CSS}{\hat{\sigma}_{CSS}} \quad (62)$$

5 Co-jump testing

While univariate jump tests have been researched extensively, the study of co-jump tests has started growing only recently. One branch of literature proposes co-jump tests through identifying jumps in a portfolio. For example, Bollerslev et al. (2008) use observed return product to construct a test statistic

for detecting co-jumps in an equi-weighted index constructed from 40 stocks. Their co-jump test detects the modest-sized common jumps ignored in the ([Barndorff-Nielsen and Shephard, 2004](#)) jump test approach. Another branch uses univariate jump tests to identify co-jump among multivariate process. For example, [Gilder et al. \(2014\)](#) propose a co-exceedance rule to identify co-jumps by using univariate jump tests. Their Monte Carlo results show that the co-exceedance rule has similar power to the co-jump test proposed by [Bollerslev et al. \(2008\)](#). The third strand develops co-jump tests which can be directly applied to multiple price processes (see, e.g., [Jacod and Todorov, 2009](#), [Bandi and Reno, 2016](#), [Bibinger and Winkelmann, 2015](#), [Caporin et al., 2017](#)). [Jacod and Todorov \(2009\)](#) propose co-jump tests based on two null hypotheses: (i) there are common jumps in a bivariate process; (ii) there are disjoint jumps in a bivariate process. [Mancini and Gobbi \(2012\)](#) construct threshold estimators for integrated covariation from the realized covariation and show that the central limit theorem and robustness to nonsynchronous data still hold under different scenarios. [Gnabo et al. \(2014\)](#) propose a co-jump test based on bootstrapping methods. [Bandi and Reno \(2016\)](#) develop a nonparametric infinitesimal moments method to detect co-jumps between asset returns and volatilities. [Bibinger and Winkelmann \(2015\)](#) propose a spectral estimation method to detect co-jumps in multivariate high-frequency data in the presence of market microstructure noise and asynchronous observations. [Caporin et al. \(2017\)](#) build a co-jump test on the comparison between smoothed realized variance and smoothed random realized variation. More related literature about co-jumps can also be seen in [Lahaye et al. \(2011\)](#) and [Dungey et al. \(2011\)](#). In the following section, we discuss five most widely used co-jump tests in details.^c

5.1 BLT co-jump testing

[Bollerslev et al. \(2008\)](#) propose a mcp test to detect co-jumps in a large ensemble of stocks. They develop a theoretical foundation which shows how only co-jumps (not idiosyncratic jumps) can be detected in a large equi-weighted index. Let n denotes the total number of assets under co-jump detection. The mcp mean cross-product test statistic is defined as:

$$mcp_{t,j} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{l=i+1}^n \Delta_j X^i \Delta_j X^l, \quad j = 1, \dots, M-1, t = 1, \dots, T \quad (63)$$

where

$$\Delta_j X^i = X_{t+(j+1)/M}^i - X_{t+j/M}^i, \quad \text{for } i = 1, \dots, n \quad (64)$$

^cWe follow the notation and description as in [Peng and Swanson \(2018\)](#).

Since the mcp-statistic has nonzero mean and is analogous to a U -statistic, the studentized test statistic is

$$z_{mcp,t,j} = \frac{mcp_{t,j} - \overline{mcp}_t}{s_{mcp,t}}, \text{ for } j = 1, \dots, M-1 \text{ and } t = 1, \dots, T. \quad (65)$$

where

$$\overline{mcp}_t = \frac{1}{M-1} mcp_t = \frac{1}{M-1} \sum_{j=1}^{M-1} mcp_{t,j} \quad (66)$$

and

$$s_{mcp,t} = \sqrt{\frac{1}{M-1} \sum_{j=1}^{M-1} (mcp_{t,j} - \overline{mcp}_t)^2} \quad (67)$$

The null distribution under the null hypothesis of no jump is derived from bootstrapping the test statistics $z_{mcp,t,j}$ under Monte Carlo simulations.

5.2 JT co-jump testing

Jacod and Todorov (2009) construct two test statistics to identify co-jumps under two different null hypothesis: (i) there is at least one common jump under the null hypothesis; (ii) there is at least one disjoint jump under the null hypothesis. The test statistics are proposed for detecting co-jumps on bivariate processes for the path of $s \rightarrow X_s$ on $[0, t]$. Co-jumps among multivariate processes can be detected from the combination of bivariate processes. The test statistics of the common jump $\Phi_n^{(j)}$ and disjoint jump $\Phi_n^{(d)}$ are defined as:

$$\Phi_n^{(j)} = \frac{V(f, k\Delta_n)_t}{V(f, \Delta_n)_t} \quad (68)$$

$$\Phi_n^{(d)} = \frac{V(f, \Delta_n)_t}{\sqrt{V(g_1, \Delta_n)_t V(g_2, \Delta_n)_t}} \quad (69)$$

where k is an integer greater than 1, and $\Delta_n = \frac{t}{M}$ is the length of equi-spaced intra-daily time interval. $V(f, k\Delta_n)_t$ is defined as:

$$V(f, k\Delta_n)_t = \sum_{j=1}^{\lfloor t/k\Delta_n \rfloor} f(X_{(j+1)k/M} - X_{jk/M}) \quad (70)$$

where the functions for $f(x)$, $g_1(x)$, and $g_2(x)$ are defined as:

$$f(x) = (x_1 x_2)^2, g_1(x) = (x_1)^4, g_2(x) = (x_2)^4 \quad (71)$$

They propose asymptotic properties and central limit theorems of these two test statistics when the mesh Δ_n approaches 0. They show that the test

statistics for the null hypothesis with disjoint jumps $\Phi_n^{(d)}$ converges stably in law to 0 on $\Omega_T^{(d)}$ and the null hypothesis with common jumps $\Phi_n^{(j)}$ converges stably in law to 1 on $\Omega_T^{(j)}$. Here $\Omega_T^{(j)}$ and $\Omega_T^{(d)}$ are defined as:

$$\Omega_T^{(j)} = \{\omega : \text{on } [0, t] \text{ the process } \Delta_j X^1 \Delta_j X^2 \text{ is not identically 0}\} \quad (72)$$

$$\begin{aligned} \Omega_T^{(j)} &= \{\omega : \text{on } [0, t] \text{ the processes } \Delta_j X^1 \text{ and } \Delta_j X^2 \text{ are} \\ &\text{not identically 0, but process } \Delta_j X^1 \Delta_j X^2 \text{ is}\} \end{aligned} \quad (73)$$

where $\Delta_j X^i = X_{(j+1)/M}^i - X_{j/M}^i$, for $i = 1, 2$ and $j = 1, \dots, M - 1$. The authors construct critical regions of the two statistics as:

$$C_n^{(j)} = \{|\Phi_n^{(j)} - 1| \geq c_n^{(j)}\} \quad (74)$$

$$C_n^{(d)} = \{\Phi_n^{(d)} \geq c_n^{(d)}\} \quad (75)$$

5.3 MG threshold co-jump test

[Mancini and Gobbi \(2012\)](#) use a threshold r_h to estimate each co-jump as:

$$\Delta_j X^1 \Delta_j X^2 - \Delta_j X^1 1_{\{(\Delta_j X^1)^2 \leq r_h\}} \Delta_j X^2 1_{\{(\Delta_j X^2)^2 \leq r_h\}}, \quad (76)$$

where h is the length of observations interval and $h = \frac{t}{M}$ for every $j = 1, \dots, M$. Threshold r_h is defined by a deterministic function from $h \rightarrow r_h$, with the following properties:

$$\lim_{h \rightarrow 0} r_h = 0 \text{ and } \lim_{h \rightarrow 0} \left(h \log \frac{1}{h} \right) / r_h = 0.$$

The threshold r_h depends on an unknown realized instantaneous volatility path. Monte Carlo simulations are used under different models to select a reasonable threshold. For example, in the model of stochastic volatility and finite compound Poisson jump part, the optimal choice of threshold is $r_h = 0.33 \widehat{IC}_{t,M} h^{0.99}$, where the integrated covariation estimator $\widehat{IC}_{t,M}$ is derived by [Mancini \(2001\)](#):

$$\widehat{IC}_{t,M} = \tilde{v}_{1,1}^{(M)}(X^1, X^2)_t, \quad (77)$$

where

$$\tilde{v}_{1,1}^{(M)}(X^1, X^2)_t = h^{-3} \sum_{j:t_j \leq t} \Delta_j X^1 1_{\{(\Delta_j X^1)^2 \leq r_h\}} \Delta_j X^2 1_{\{(\Delta_j X^2)^2 \leq r_h\}}, \quad (78)$$

5.4 GST co-exceedance rule

[Gilder et al. \(2014\)](#) propose a co-exceedance based co-jump detection method by applying univariate jump tests to individual stocks to identify co-jumps.

They select three univariate jump tests in [Barndorff-Nielsen and Shephard \(2006\)](#), [Lee and Mykland \(2007\)](#), and [Andersen et al. \(2010\)](#). The co-jumps are detected as intersection between ABD jump test results and BNS jump test results ($ABD \cap BNS$), intersection between ABD jump test results and LM jump test results ($ABD \cap LM$), intersection between BNS jump test results and LM jump test results ($BNS \cap LM$), and the intersection among three jump tests results ($ABD \cap LM \cap BNS$).

The nonparametric BNS jump test and LM jump test have been discussed in [Sections 4.1](#) and [4.2](#), respectively. The ABD jump test in [Andersen et al. \(2010\)](#) is the sequential BNS test which first identifies jump days through BNS test and then calculates the maximum intra-day return as the jump level. [Gilder et al. \(2014\)](#) modified the maximum intra-day return during jump days into:

$$\max(|\Delta_j X| / \sqrt{s_{WSD,j}^2 \cdot \Delta \cdot BPV_t}), \text{ for } j = 1, \dots, M-1 \quad (79)$$

where $\Delta_j X = X_{t+(j+1)/M} - X_{t+j/M}$ for $t = 1, \dots, T$ and $\Delta = \frac{1}{M}$. Here $s_{WSD,j}^2$ is the weighted standard deviation (WSD) estimator proposed by [Boudt et al. \(2011\)](#).

Comparisons between co-exceedance rule for co-jump detection and BLT co-jump test are made under extensive Monte Carlo simulations. The results show that intra-day co-exceedance-based detection method has similar power to that of the BLT co-jump test both on large and small co-jumps.

5.5 CKR co-jump testing

The test statistics in [Caporin et al. \(2017\)](#) is derived from the difference between smoothed realized variance (SRV) and smoothed randomized realized variance ($SRRV$). The SRV is denoted as:

$$SRV(X^i) = \sum_{j=1}^M |\Delta_j X^i|^2 \cdot K\left(\frac{\Delta_j X^i}{H_{\Delta_j, M}^i}\right) \cdot \eta_j^i, \quad i = 1, \dots, n, \quad (80)$$

where $K(\cdot)$ is a differentiable kernel function with bounded first derivative almost everywhere in R having the following properties:

$$K(0) = 1, 0 \leq K(\cdot) \leq 1, \text{ and } \lim_{x \rightarrow \infty} K(|x|) = 0 \quad (81)$$

And H is the bandwidth which is denoted as:

$$H_{\Delta_j, M}^i = h_M \cdot \widehat{\sigma}_{\Delta_j}^i \sqrt{\frac{t}{M}} \quad (82)$$

where h_M is the bandwidth parameter and $\widehat{\sigma}_{\Delta_j}^i$ is the point estimator of the local standard deviation of i th asset. η_j^i is an $n \times M$ matrix independent and identically distributed variable such that $E[\eta_j^i] = 1$ and $Var[\eta_j^i] = V_\eta \leq \infty$. V_η is set to 0.0025 in the application of the test.

Another estimator (\widetilde{SRV}) is written in the form as:

$$\widetilde{SRV}^n(X^i) = \sum_{j=1}^M |\Delta_j X^i|^2 \cdot \left(K \left(\frac{\Delta_j X^i}{H_{\Delta_j, M}^i} \right) + \pi_{k=1}^n \left(1 - K \left(\frac{\Delta_j X^k}{H_{\Delta_j, M}^k} \right) \right) \right) \quad (83)$$

The proposed test statistics takes the form:

$$S_{M,n} = \frac{1}{V_n} \sum_{i=1}^M \frac{(SRRV(X^i) - \widetilde{SRV}^n(X^i))^2}{SQ(X^i)}, \quad (84)$$

where

$$SQ(X^i) = \sum_{j=1}^M |\Delta_j X^i|^4 \cdot K^2 \left(\frac{\Delta_j X^i}{H_{\Delta_j, M}^i} \right), \quad i = 1, \dots, n \quad (85)$$

The asymptotic behavior of the $S_{n,N}$ is described as:

$$S_{M,n} \xrightarrow{d} \chi^2(n), \quad \text{on } \bar{\Omega}_T^n$$

$$S_{M,n} \xrightarrow{P} +\infty, \quad \text{on } \bar{\Omega}_T^{MJ,n}$$

where $\bar{\Omega}_T^n$ and $\bar{\Omega}_T^{MJ,n}$ is defined as:

$$\begin{aligned} \bar{\Omega}_T^{MJ,n} &= \{ \omega \in \Omega \mid \prod_{i=1}^n (\Delta X^i)_t \text{ is not identically 0} \}, \\ \bar{\Omega}_T^n &= \Omega / \bar{\Omega}_T^{MJ,n} \end{aligned}$$

6 Empirical experiments

6.1 Data description

The empirical experiments are conducted with six stocks and two ETFs. The six individual stocks, which include the Boeing Company (BA), Exxon Mobile Corporation (XOM), Johnson & Johnson (JNJ), JPMorgan Chase & Co. (JPM), Microsoft Corporation (MSFT), and Walmart Inc.(WMT), have the highest weight in their corresponding SPDR market sector ETFs such as XLI (industrial sector), XLE (energy sector), XLV (healthcare sector), XLF (finance sector), XLK (technology sector), and XLP (consumer staples sector). The two SPDR sector ETFs chosen are the energy and technology sector ETFs and XLE & XLK. The dataset is obtained from the Trade and Quote Database (TAQ) of Wharton Research Data Service (WRDS) and it covers the period from January 1, 2006 to December 31, 2013 for a total of 2013 days. We select trade data ranging from 9:30 am to 4 pm on regular trading days. Overnight transactions are excluded from our dataset. We mainly use a 5-min sampling frequency to eradicate the effect of market microstructure noise in the data which yields 78 total observations per day. We also use a 1-min sampling frequency in

specific cases which yields 390 observations per day. It should be noted that all empirical experiments are carried out on the logarithmic values of the stock and ETF prices.

6.2 Methodology

Our empirical experiment consists of three sections: (i) integrated volatility measures, (ii) jump tests, and (iii) co-jump tests. For each of the different parts, we conduct analysis involving the most widely used measures and tests, respectively. A detailed description of the different measures and tests used and the empirical methodologies thereof is given as follows.

First, we use six different measures to estimate Integrated Volatility for all the stocks and ETFs: (1) RV ([Section 3.1](#)), (2) BPV ([Section 3.2](#)), (3) TPV ([Section 3.3](#)), (4) TRV ([Section 3.7](#)), (5) MedRV, and (6) MinRV ([Section 3.11](#)). Second, to test for price jumps in the data three different jump tests are used: (1) ASJ jump test ([Section 4.4](#)), (2) BNS jump test ([Section 4.1](#)), and (3) LM jump test ([Section 4.2](#)). Lastly, co-jump tests are carried out using (1) JT co-jump test ([Section 5.2](#)), (2) BLT co-jump test ([Section 5.1](#)), and (3) GST co-exceedance rule ([Section 5.4](#)).

Estimation of integrated volatility, BNS and LM jump tests as well as all the co-jump tests are carried out using 5-min data where Δ is set to $\frac{1}{78}$. However, for the ASJ jump test, both $1 - (\Delta = \frac{1}{390})$ and 5-min frequencies are used as a basis for comparative study.

When conducting analysis using jump tests, we calculate the percentage of days identified as having jumps. For both the BNS and ASJ tests, it can be given as:

$$\text{Percentage of jump days} = \frac{100 \sum_{i=1}^T I(Z_i > c_\alpha)}{T} \% \quad (86)$$

where $I(\cdot)$ is the jump indicator function, c_α is the critical value at α significance level and Z_i is the BNS or ASJ jump test statistics. For the LM jump test on the other hand it can be derived as:

$$\text{Percentage of jump days} = \frac{100 \sum_{i=0}^T I(\exists t \in i, |L_t| > c_\alpha)}{T} \% \quad (87)$$

where L_t is the LM jump test statistic at the intra-day level within a particular day, t refers to the 78 intra-day intervals and c_α is the critical value at α significance level.

Once jumps are detected, we follow [Andersen et al. \(2007\)](#) and [Duong and Swanson \(2011\)](#) to construct risk measures by separating out the variation due to daily jump component and the continuous components. This is done by using volatility measures *RV* and *TPV*. It can be given as:

$$\text{Variation due to jump component} = JV_t = \max[RV_t - TPV_t, 0] * I_{jump,t} \quad (88)$$

Consequently the ratio of jump to total variation for all three jump tests can be calculated as:

$$\text{Ratio of jump variation to total variation} = \frac{JV_t}{RV_t} \quad (89)$$

For BLT co-jump test, the percentage of days identified as having co-jumps is calculated using:

$$\text{Percentage of co-jump days} = \frac{100 \sum_{i=0}^T I(\exists j, z_{mcp,i,j} < c_{mcp,\alpha,l} \cup z_{mcp,i,j} > c_{mcp,\alpha,r})}{T} \% \quad (90)$$

where $c_{mcp,\alpha,l}$ and $c_{mcp,\alpha,r}$ are left and right tail critical values derived from bootstrapping the null distribution. α is the significance level. For the JT co-jump test, the percentage of days identified as having co-jumps is calculated as:

$$\text{Percentage of co-jump days} = \frac{100 \sum_{i=0}^T I(\Phi_n^{(d)} \geq c_n^{(d)})}{T} \% \quad (91)$$

In the co-exceedance rule proposed by [Gilder et al. \(2014\)](#), we use the BNS jump test and the LM jump test to identify co-jumps. The percentage of days identified as having co-jumps can be given as:

$$\text{Percentage of co-jump days} = \frac{100 \sum_{i=0}^T I(|Z_i| \geq \Phi_\alpha) * I(\exists t \in i, |L_t| > c_\alpha)}{T} \% \quad (92)$$

where Z_i is the BNS jump test statistic and L_t is the LM jump test statistic.

In addition to reporting the findings of our empirical experiment on the entire sample, we also conduct analysis after splitting the data set into two periods. The first sample consists of the period from January 2006 to June 2009 and the second sample consists of the period from July 2009 to December 2012. This is done to inspect whether the jump activity in the stocks and the ETFs changes considerably over time. The break date of our sample (June 2009) roughly corresponds to the end of the business cycle contraction after the financial crisis as given by NBER.

6.3 Findings

[Table 1](#) gives the summary statistics for integrated volatility which is estimated using six volatility measures RV , BPV , TPV , $MedRV$, $MinRV$, and TRV . The sample period considered for the six stocks and the two ETFs is

TABLE 1 Descriptive statistics of integrated volatility measures: sample period January 2006–December 2013

| | Volatility measures | RV | BPV | TPV | MedRV | MinRV | TRV |
|-------------------|---------------------|--------|--------|--------|--------|--------|--------|
| Boeing | Mean | 2.68 | 1.24 | 2.29 | 2.38 | 2.46 | 2.60 |
| | Standard deviation | 4.35 | 2.10 | 3.95 | 4.10 | 4.29 | 4.32 |
| | Min | 0.21 | 0.10 | 0.15 | 0.19 | 0.20 | 0.21 |
| | Max | 60.05 | 28.47 | 58.80 | 62.83 | 65.26 | 60.05 |
| | Skewness | 6.50 | 6.66 | 6.81 | 7.13 | 7.00 | 6.63 |
| | Kurtosis | 61.54 | 62.81 | 66.26 | 73.91 | 70.90 | 63.35 |
| Exxon | Mean | 2.15 | 1.01 | 1.89 | 1.98 | 2.02 | 2.10 |
| | Standard deviation | 5.30 | 2.65 | 4.94 | 5.47 | 5.75 | 5.08 |
| | Min | 0.10 | 0.04 | 0.07 | 0.08 | 0.09 | 0.10 |
| | Max | 131.00 | 72.39 | 140.95 | 156.49 | 166.62 | 131.00 |
| | Skewness | 12.17 | 14.05 | 15.01 | 15.47 | 15.91 | 12.89 |
| | Kurtosis | 223.33 | 301.42 | 346.86 | 357.84 | 374.71 | 254.66 |
| Johnson & Johnson | Mean | 1.05 | 0.47 | 0.85 | 0.90 | 0.93 | 1.00 |
| | Standard deviation | 2.37 | 1.06 | 1.94 | 2.07 | 2.12 | 2.30 |
| | Min | 0.07 | 0.02 | 0.03 | 0.05 | 0.05 | 0.07 |
| | Max | 52.86 | 22.84 | 46.48 | 42.82 | 43.33 | 52.86 |
| | Skewness | 11.36 | 10.44 | 11.25 | 10.99 | 10.61 | 11.76 |
| | Kurtosis | 186.83 | 157.67 | 195.54 | 173.99 | 162.96 | 203.20 |

Continued

TABLE 1 Descriptive statistics of integrated volatility measures: sample period January 2006–December 2013—Cont'd

| | Volatility measures | RV | BPV | TPV | MedRV | MinRV | TRV |
|-----------|---------------------|--------|--------|--------|--------|--------|--------|
| JPMorgan | Mean | 5.85 | 2.76 | 5.11 | 5.26 | 5.41 | 5.80 |
| | Standard deviation | 13.99 | 65.23 | 12.00 | 12.24 | 12.49 | 13.96 |
| | Min | 0.13 | 0.05 | 0.09 | 0.11 | 0.10 | 0.13 |
| | Max | 244.81 | 118.46 | 213.02 | 214.80 | 235.11 | 244.81 |
| | Skewness | 8.06 | 7.74 | 7.48 | 7.34 | 7.42 | 8.10 |
| | Kurtosis | 100.06 | 94.76 | 88.59 | 85.07 | 90.95 | 101.07 |
| Microsoft | Mean | 2.29 | 1.07 | 1.95 | 2.07 | 2.12 | 2.24 |
| | Standard deviation | 3.71 | 1.82 | 3.49 | 3.62 | 3.67 | 3.69 |
| | Min | 0.16 | 0.06 | 0.10 | 0.13 | 0.14 | 0.14 |
| | Max | 62.08 | 34.69 | 66.05 | 54.42 | 65.94 | 62.08 |
| | Skewness | 7.32 | 7.98 | 8.14 | 7.44 | 7.76 | 7.44 |
| | Kurtosis | 82.10 | 99.39 | 10.17 | 79.17 | 91.30 | 83.96 |
| Walmart | Mean | 1.57 | 0.71 | 1.31 | 1.37 | 1.40 | 1.51 |
| | Standard deviation | 2.95 | 1.32 | 2.53 | 2.53 | 2.48 | 2.86 |
| | Min | 0.13 | 0.57 | 0.10 | 0.12 | 0.11 | 0.13 |
| | Max | 71.09 | 31.41 | 63.15 | 60.95 | 53.54 | 71.09 |
| | Skewness | 10.74 | 10.07 | 10.93 | 10.19 | 8.76 | 11.02 |
| | Kurtosis | 190.00 | 172.43 | 205.05 | 180.79 | 130.01 | 204.82 |

| | | | | | | | |
|-----|--------------------|--------|--------|--------|--------|--------|--------|
| XLE | Mean | 2.82 | 1.35 | 2.49 | 2.70 | 2.74 | 2.72 |
| | Standard deviation | 7.35 | 3.45 | 5.77 | 8.52 | 8.42 | 6.00 |
| | Min | 0.10 | 0.04 | 0.08 | 0.08 | 0.08 | 0.10 |
| | Max | 193.76 | 80.53 | 123.39 | 278.12 | 266.10 | 129.66 |
| | Skewness | 14.29 | 12.30 | 10.14 | 20.21 | 18.81 | 10.39 |
| | Kurtosis | 296.35 | 215.74 | 151.36 | 577.77 | 511.09 | 161.29 |
| XLK | Mean | 1.46 | 0.65 | 1.19 | 1.28 | 1.30 | 1.38 |
| | Standard deviation | 3.18 | 1.41 | 2.72 | 2.69 | 2.78 | 2.81 |
| | Min | 0.07 | 0.02 | 0.04 | 0.06 | 0.05 | 0.07 |
| | Max | 61.21 | 24.11 | 49.50 | 40.78 | 41.31 | 49.62 |
| | Skewness | 8.88 | 7.92 | 8.53 | 7.52 | 7.72 | 7.83 |
| | Kurtosis | 117.31 | 87.29 | 103.04 | 76.99 | 79.80 | 89.51 |

Notes: Table 1 gives the descriptive statistics of the different integrated volatility measures. Mean, standard deviation, and min and max values are all in terms of 10^{-4} .

January 2006–December 2013. The mean, standard deviation, minimum, and maximum values are all in terms of 10^{-4} . Among all the stocks and ETFs, JPMorgan seems to have undergone maximum price fluctuations across the sample period as it displays the highest mean and max values across all the volatility measures. On the other hand Johnson & Johnson and XLK appear to be tied in terms of having undergone least amount of price fluctuations as they display the lowest mean and max volatility estimates. Among all the volatility measures, *BPV* reports the lowest mean volatility estimate while *RV* reports the highest mean volatility estimate for any given stock or ETF. This can be explained by the fact that in the presence of frequent jumps, *RV* overestimates integrated volatility. To get a clearer idea of how volatility differs across the stocks and ETFs, we turn to Figs. 1 and 2 which display the estimated volatility for the stocks Boeing and Exxon with respect to the six aforementioned volatility measures. Similar figures for four other stocks and two ETFs have not been given for the purpose of brevity and can be provided upon request. In general stocks and ETFs achieve their highest volatility in the

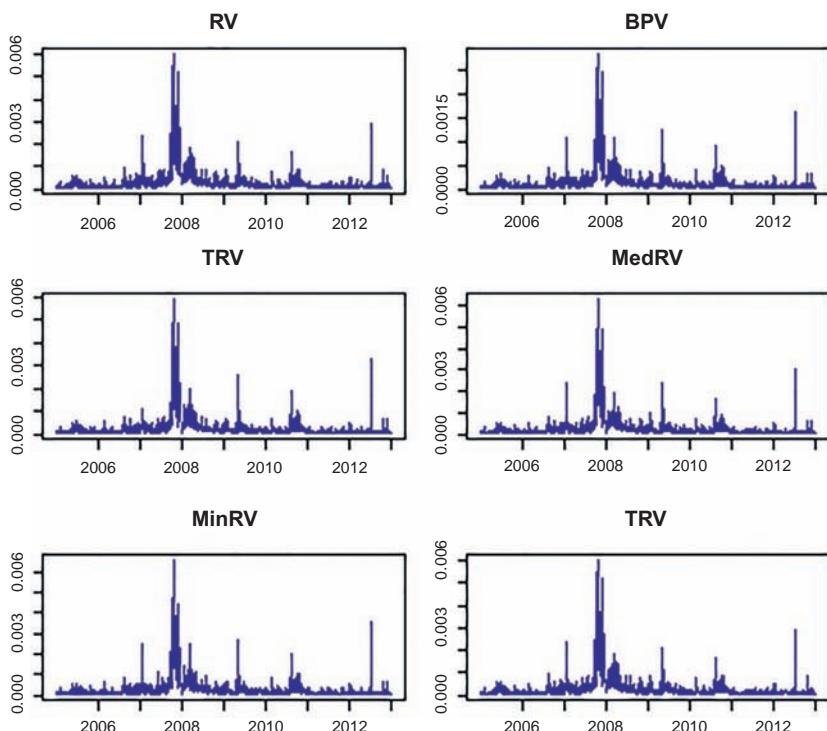


FIG. 1 Integrated volatility measures—Boeing. Notes: Fig. 1 displays volatility of Boeing across the sample period January 2006–December 2013 using 5-min sampling frequency with respect to six different integrated volatility measures which include *RV*, *BPV*, *TPV*, *MedRV*, *MinRV*, and *TRV*.

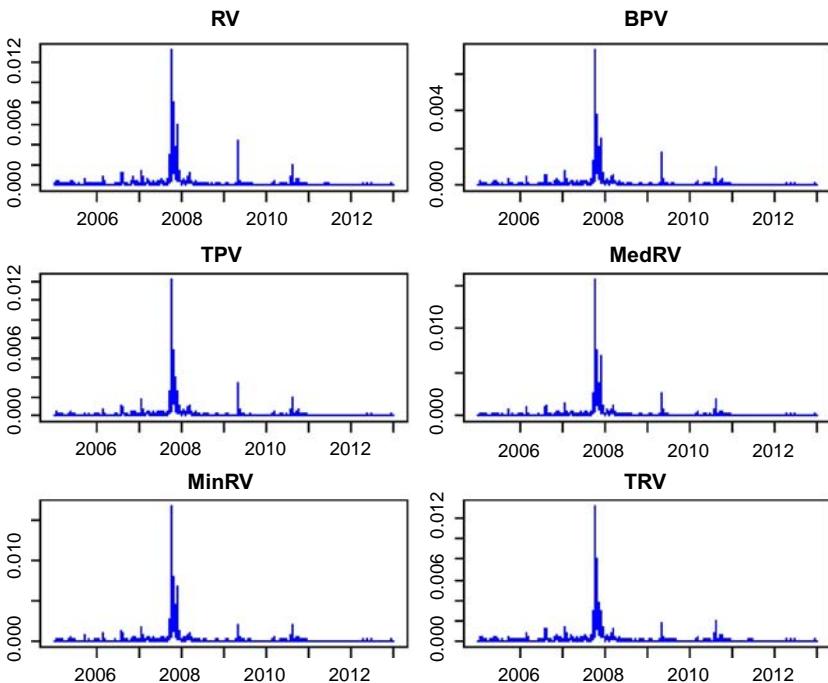


FIG. 2 Integrated volatility measures—Exxon. Notes: Fig. 2 displays volatility movement of Exxon. See notes of Fig. 1.

fourth quarter of 2008 during the financial crisis with a few exceptions. For XLE, in case of all four volatility measures apart from *TPV* and *TRV*, volatility reaches its peak in the second quarter of 2009. For XLK on the other hand, only in case *RV* the volatility peak is reached in the first quarter of 2008 while for the other measures it is the fourth quarter of 2008.

We now look at Tables 2–5 which display the descriptive statistics of the three jump tests. For the ASJ jump test we consider both 5- and 1-min frequencies while for the BNS and the LM jump tests we only consider 5-min frequency. Panel A in the tables refers to the prefinancial crisis sample period, January 2006–June 2009 and panel B refers to the postcrisis period July 2009–December 2012. In case of the ASJ jump tests, we find noticeable differences between 5- (Table 2) and 1-min (Table 3) frequencies. Overall the mean value of the statistics is higher for the 1-min data compared to the 5-min frequency suggesting that more jumps would be identified in the 1-min case. The skewness values are all negative irrespective of the sample period, type of stock and frequency of sampling suggesting that the ASJ test statistics are left-skewed. Panel A for both frequencies appear to have overall higher mean and max values again suggesting more jump activity in the financial crises period. In case of the BNS test (Table 4) the skewness values are all positive,

TABLE 2 Descriptive statistics of ASJ jump test: 5-min sampling frequency

| | | Boeing | Exxon | Johnson & Johnson | JPMorgan | Microsoft | Walmart | XLE | XLK |
|---------|--------------------|--------|--------|-------------------|----------|-----------|---------|--------|--------|
| Panel A | Mean | 0.042 | 0.177 | 0.256 | 0.088 | 0.149 | 0.185 | 0.016 | 0.164 |
| | Standard deviation | 1.239 | 1.118 | 1.166 | 1.229 | 1.189 | 1.188 | 1.243 | 1.113 |
| | Skewness | -1.129 | -1.149 | -1.293 | -1.236 | -1.150 | -1.340 | -1.182 | -1.134 |
| | Kurtosis | 4.120 | 4.459 | 4.739 | 4.814 | 4.253 | 4.752 | 4.612 | 4.583 |
| | Max | 2.413 | 2.648 | 2.330 | 2.427 | 2.530 | 2.666 | 2.938 | 2.526 |
| | Min | -5.374 | -4.383 | -4.681 | -5.833 | -4.319 | -4.363 | -5.746 | -4.361 |
| Panel B | Mean | -0.053 | 0.093 | -0.009 | -0.056 | 0.109 | 0.0541 | -0.018 | 0.063 |
| | Standard deviation | 1.291 | 1.168 | 1.302 | 1.273 | 1.182 | 1.146 | 1.192 | 1.158 |
| | Skewness | -1.074 | -1.216 | -1.211 | -1.009 | -1.079 | -0.840 | -1.150 | -1.195 |
| | Kurtosis | 3.743 | 4.465 | 4.342 | 3.636 | 4.046 | 3.402 | 4.535 | 4.808 |
| | Max | 2.210 | 2.301 | 2.431 | 2.300 | 2.563 | 2.367 | 2.783 | 2.495 |
| | Min | -5.095 | -4.495 | -5.685 | -4.624 | -4.437 | -3.963 | -5.770 | -5.658 |

Notes: Table 2 gives the descriptive statistics of ASJ jump test at 5-min frequency. Panel A covers the financial crisis period from January 2006 to June 2009 and panel B covers the postfinancial crisis period from July 2009 to December 2012.

TABLE 3 Descriptive statistics of ASJ jump test: 1-min sampling frequency

| | | Boeing | Exxon | Johnson & Johnson | JPMorgan | Microsoft | Walmart | XLE | XLK |
|---------|--------------------|--------|--------|-------------------|----------|-----------|---------|--------|--------|
| Panel A | Mean | 0.399 | 0.404 | 0.587 | 0.391 | 0.612 | 0.453 | 0.125 | 0.756 |
| | Standard deviation | 1.455 | 1.345 | 1.414 | 1.315 | 1.300 | 1.358 | 1.371 | 1.385 |
| | Skewness | -0.896 | -0.776 | -1.627 | -1.035 | -1.05 | -0.845 | -0.347 | -1.071 |
| | Kurtosis | 4.378 | 5.219 | 10.193 | 6.531 | 5.650 | 4.770 | 3.682 | 9.357 |
| | Max | 4.723 | 5.138 | 4.538 | 4.105 | 5.457 | 4.998 | 4.548 | 6.183 |
| | Min | -6.235 | -6.387 | -11.076 | -8.702 | -5.554 | -5.755 | -4.864 | -8.529 |
| Panel B | Mean | 0.236 | 0.348 | 0.430 | 0.249 | 0.532 | 0.317 | 0.066 | 0.643 |
| | Standard deviation | 1.400 | 1.281 | 1.504 | 1.335 | 1.247 | 1.503 | 1.321 | 1.111 |
| | skewness | -1.100 | -1.122 | -0.884 | -1.294 | -1.029 | -1.279 | -0.711 | -0.568 |
| | Kurtosis | 5.678 | 7.305 | 4.546 | 7.522 | 5.3471 | 7.099 | 3.819 | 4.707 |
| | Max | 4.244 | 4.867 | 5.120 | 4.198 | 4.684 | 3.505 | 4.265 | 5.429 |
| | Min | -8.733 | -9.165 | -6.190 | -9.994 | -5.867 | -10.974 | -4.813 | -3.816 |

Notes: Table 3 gives the descriptive statistics of ASJ jump test at 1-min frequency. See notes of Table 9.

TABLE 4 Descriptive statistics of BNS jump test

| | | Boeing | Exxon | Johnson & Johnson | JPMorgan | Microsoft | Walmart | XLE | XLK |
|---------|--------------------|--------|--------|-------------------|----------|-----------|---------|--------|--------|
| Panel A | Mean | 0.878 | 0.771 | 1.098 | 0.854 | 0.867 | 0.859 | 0.599 | 1.415 |
| | Standard deviation | 1.511 | 1.327 | 1.665 | 1.434 | 1.320 | 1.401 | 1.327 | 1.693 |
| | Skewness | 1.030 | 0.618 | 1.151 | 0.951 | 0.602 | 0.801 | 0.780 | 1.315 |
| | Kurtosis | 4.903 | 3.749 | 4.851 | 4.518 | 3.500 | 4.068 | 4.169 | 6.395 |
| | Max | 7.791 | 6.631 | 8.661 | 7.352 | 5.811 | 6.662 | 6.947 | 10.731 |
| | Min | -2.513 | -2.341 | -2.693 | -2.364 | -2.299 | -2.505 | -2.640 | -2.695 |
| Panel B | Mean | 0.688 | 0.667 | 0.876 | 0.558 | 0.880 | 0.918 | 0.572 | 0.814 |
| | Standard deviation | 1.406 | 1.328 | 1.450 | 1.249 | 1.406 | 1.535 | 1.228 | 1.275 |
| | Skewness | 0.983 | 0.572 | 0.837 | 0.574 | 0.647 | 0.925 | 0.615 | 0.525 |
| | Kurtosis | 5.400 | 3.557 | 3.646 | 3.652 | 3.428 | 4.215 | 3.711 | 3.530 |
| | Max | 9.340 | 6.964 | 7.128 | 5.486 | 7.008 | 7.549 | 5.807 | 6.909 |
| | Min | -2.485 | -2.546 | -2.006 | -2.750 | -2.417 | -2.137 | -2.225 | -2.244 |

Notes: Table 4 gives the descriptive statistics of BNS jump test at 5-min frequency. See notes of Table 2.

TABLE 5 Descriptive statistics of LM jump test

| | | Boeing | Exxon | Johnson & Johnson | JPMorgan | Microsoft | Walmart | XLE | XLK |
|---------|--------------------|-----------|----------|-------------------|-----------|-----------|-----------|----------|----------|
| Panel A | Mean | 0.595 | 4.436 | -1.662 | -0.676 | 3.987 | 0.181 | 2.202 | 2.347 |
| | Standard deviation | 42.865 | 50.142 | 101.707 | 47.879 | 66.636 | 52.002 | 37.687 | 72.997 |
| | Skewness | 0.911 | 2.909 | 0.675 | 0.850 | 1.267 | -0.146 | 2.714 | 0.316 |
| | Kurtosis | 15.591 | 31.900 | 16.875 | 25.186 | 13.907 | 10.491 | 41.367 | 14.882 |
| | Max | 386.170 | 602.678 | 1008.160 | 525.995 | 585.419 | 314.852 | 472.919 | 509.147 |
| | Min | -309.16 | -185.941 | -615.524 | -354.424 | -289.795 | -343.618 | -189.563 | -570.747 |
| Panel B | Mean | -0.110 | 0.295 | 0.076 | 0.137 | 0.097 | 0.636 | -0.051 | 0.042 |
| | Standard deviation | 42.450 | 47.158 | 67.736 | 33.483 | 42.875 | 61.795 | 43.314 | 54.955 |
| | Skewness | -0.465 | -0.251 | -0.635 | 1.958 | 1.035 | 0.138 | -0.451 | 0.128 |
| | Kurtosis | 67.346 | 29.672 | 49.711 | 165.139 | 92.97 | 76.991 | 48.134 | 30.865 |
| | Max | 1047.900 | 1192.159 | 1455.508 | 1320.357 | 1726.992 | 1529.471 | 1288.848 | 1571.141 |
| | Min | -1275.319 | -764.818 | -1789.208 | -1399.492 | -1064.302 | -2124.609 | -881.903 | -881.418 |

Notes: Table 5 gives the descriptive statistics of LM jump test at 5-min frequency. See notes of Table 2.

which suggests all BNS test statistics are right-skewed and have a long right tail. The kurtosis values are all above 3, which indicates the empirical distribution of BNS test statistics is leptokurtic. For the LM test (Table 5), a window size of $k = 50$ is chosen. The mean of LM test statistics is around 0, while the max and min value of test statistics are far from 0, even reaching 1726.992 and -2124.609.

Tables 6–8 denote the percentage of days identified as having jumps for the ASJ, BNS, and LM jump tests. For all jump tests $\alpha = 0.1$ and 0.05 significance levels are considered. In case of the ASJ jump test (Table 6), it appears that Johnson & Johnson has the largest percentage of jump days for post-June 2009 period (panel B). However, for the pre-June 2009 period (panel A), only with 5-min frequency, Johnson & Johnson attains the highest jump day percentage. While for 1-min frequency XLK seems to lead the race. XLE on the other hand has the lowest percentage of jump days across all significance levels, sample periods and sampling frequencies. In case of the BNS jump test (Table 7) XLK has the largest percentage of jump days for the crisis (panel A) period while Microsoft displays the highest percentage in the postcrisis (panel B) period. Overall for all the stocks and ETFs for both ASJ and BNS tests, panel A displays relatively higher jump activity than panel B which shows that jumps happen more frequently during financial crisis period, when compared with postfinancial crisis period. Table 8 shows the percentage of jump days and jump proportions for the LM test. The percentage of jump days is very large, reaching 90% in some cases. This is because LM jump test detects whether there is jump at each interval per day^d and a day is classified as a jump day if a jump occurs on any of the 5-min (78 observations) intervals. The jump proportion is calculated by the total number of test statistics which indicate jumps divided by the total number of test statistics across all intra-day intervals for the entire sample period. The jump proportions are much lower, close to 1%. It is noteworthy that both percentage of jump days and jump proportions are larger during postfinancial crisis period (panel B). One reason for this may be, the fact that the LM test detects more small and moderate jumps when compared with the ASJ and BNS tests and these types of jumps are more likely to happen during post-financial crisis period.

To graphically illustrate the level of jump activity we turn to Figs. 3–5, which display the ASJ and BNS test statistic values for the days identified as having jumps for Boeing and Exxon across the sample period January 2006–December 2013. Once again similar figures for other stocks and ETFs have not been given for the purpose of brevity and would be available upon request. The significance level considered is 5%. For the ASJ test, the analysis is carried out for both 5- (Fig. 3) and 1-min (Fig. 4) frequencies. As is evident from the figures, with a higher sampling frequency of 1 min, more jumps are

^dRefer to Eq. (92).

TABLE 6 Percentage of days identified as having jumps—ASJ jump test

| Name | Panel A: January 2006–June 2009 | | | | Panel B: July 2009–December 2012 | | | |
|--------------------|---------------------------------|-------|-------|-------|----------------------------------|-------|-------|-------|
| | 1 min | | 5 min | | 1 min | | 5 min | |
| Significance level | 0.05 | 0.10 | 0.05 | 0.10 | 0.05 | 0.10 | 0.05 | 0.10 |
| Boeing | 15.79 | 27.95 | 3.29 | 10.79 | 10.89 | 23.38 | 2.27 | 9.98 |
| Exxon | 12.61 | 22.84 | 3.63 | 11.47 | 9.76 | 21.11 | 2.72 | 8.74 |
| Johnson & Johnson | 19.88 | 32.04 | 4.77 | 14.09 | 16.00 | 27.80 | 3.40 | 10.44 |
| JPMorgan | 12.38 | 24.20 | 3.52 | 11.36 | 10.32 | 18.84 | 3.06 | 8.05 |
| Microsoft | 17.04 | 31.70 | 3.52 | 11.36 | 14.18 | 25.42 | 4.08 | 11.12 |
| Walmart | 15.11 | 27.61 | 3.52 | 11.59 | 13.16 | 24.85 | 3.74 | 9.53 |
| XLE | 10.22 | 17.84 | 3.06 | 9.43 | 7.60 | 13.96 | 2.27 | 7.94 |
| XLK | 20.45 | 34.31 | 3.63 | 10.68 | 15.20 | 26.44 | 2.83 | 9.19 |

Notes: Table 6 gives the percentage of days identified as having jumps by the ASJ test at both 5- and 1-min sampling frequencies. Jumps are tested at $\alpha = 0.05$ and $\alpha = 0.1$ significance level. Percentage of days is calculated using Eq. (87). Panel A covers the financial crisis period from January 2006–June 2009 and panel B covers the postfinancial crisis period from July 2009–December 2012.

TABLE 7 Percentage of days identified as having jumps—BNS jump test

| Name | Panel A: January 2006–June 2009 | | Panel B: July 2009–December 2012 | |
|--------------------|---------------------------------|-------|----------------------------------|-------|
| Significance level | 0.05 | 0.10 | 0.05 | 0.10 |
| Boeing | 20.68 | 27.73 | 16.69 | 23.04 |
| Exxon | 17.95 | 26.25 | 15.78 | 23.27 |
| Johnson & Johnson | 24.09 | 30.23 | 20.43 | 25.88 |
| JPMorgan | 19.89 | 25.23 | 12.83 | 19.64 |
| Microsoft | 18.86 | 25.23 | 21.45 | 27.70 |
| Walmart | 19.77 | 27.38 | 20.66 | 26.67 |
| XLE | 15.11 | 21.36 | 12.71 | 19.41 |
| XLK | 30.00 | 38.18 | 17.93 | 25.54 |

Notes: Table 7 shows percentage of days identified as having jumps by BNS test at 5-min sampling frequency. See notes of Table 6.

detected across all stocks and ETFs in comparison to 5-min frequency. In case of the BNS test both XLK and Johnson & Johnson appear to have the relatively higher degree of jump activity compared to the other stocks in the pre-June 2009 period, a result which evidently aligns with what we deduced from Table 7.

Figs. 6–8 contain the kernel density plots of ASJ, BNS, and LM test statistics. In case of the ASJ test statistics (Fig. 6), it appears that the distribution is left-tailed or negatively skewed. On the other hand the underlying distribution for the BNS test statistic (Fig. 7) appears to be skewed right. The LM test statistics (Fig. 8) display a high kurtosis and a long tail. All these results are consistent with what we found from Tables 2 to 5.

When analyzing the average ratio of jump variation to total variation, we compare the results between the ASJ test, BNS test and LM test. For all three tests given in Tables 9–11, ratio of jump variation to total variation is larger during financial crisis than postfinancial crisis and this result is robust across all significance levels. The tech sector ETF XLK has the largest jump variation ratio among all stocks. The BNS jump test is more likely to detect large jumps, especially during financial crisis period which is why the jump variation ratio reported by it is larger than the other tests.

Table 12 contains percentage of days identified as having co-jumps under both JT co-jump test and the co-exceedance rule between BNS jump test

TABLE 8 Percentage of days identified as having jumps and jump proportion—LM jump test

| Name | Panel A: January 2006–June 2009 | | | | Panel B: July 2009–December 2012 | | | |
|--------------------|---------------------------------|------|----------------|-------|----------------------------------|------|----------------|-------|
| | Jump proportion | | % of jump days | | Jump proportion | | % of Jump days | |
| Significance level | 0.05 | 0.10 | 0.05 | 0.10 | 0.05 | 0.10 | 0.05 | 0.10 |
| Boeing | 1.06 | 1.07 | 82.39 | 83.41 | 1.14 | 1.15 | 88.99 | 89.56 |
| Exxon | 1.10 | 1.11 | 85.45 | 86.36 | 1.17 | 1.17 | 90.92 | 91.49 |
| Johnson & Johnson | 1.15 | 1.15 | 89.55 | 89.89 | 1.21 | 1.21 | 94.32 | 94.55 |
| JPMorgan | 1.03 | 1.04 | 80.45 | 80.91 | 1.11 | 1.12 | 86.83 | 87.29 |
| Microsoft | 1.15 | 1.15 | 89.66 | 89.66 | 1.16 | 1.17 | 90.47 | 91.15 |
| Walmart | 1.11 | 1.12 | 86.36 | 87.05 | 1.19 | 1.19 | 92.51 | 92.74 |
| XLE | 1.03 | 1.05 | 80.45 | 81.93 | 1.17 | 1.18 | 91.60 | 92.30 |
| XLK | 1.06 | 1.07 | 82.84 | 83.41 | 1.14 | 1.14 | 88.88 | 89.22 |

Notes: Table 8 shows percentage of days identified as having jumps and jump proportion as per the LM test at 5-min sampling frequency. Percentage of days is calculated using Eq. (87). See notes of Table 6.

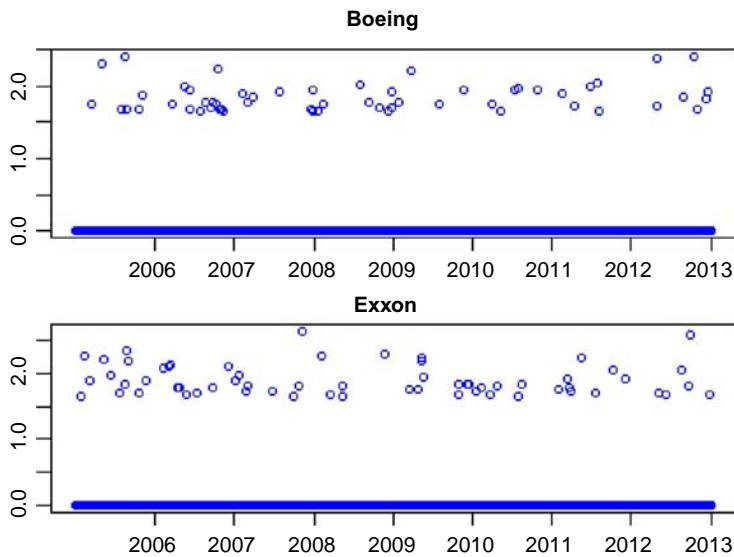


FIG. 3 ASJ jump test statistics of days identified as having jumps: 5-min sampling frequency.
Notes: Fig. 3 displays the scatter plot for ASJ test statistics for days identified as having jumps using 5-min sampling frequency. We consider the following stocks and ETFs: Boeing & Exxon for the sample period January 2006–December 2013.

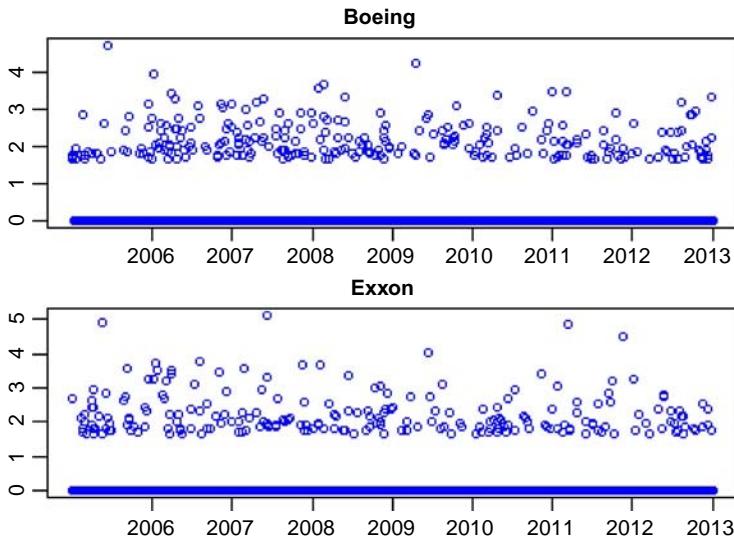


FIG. 4 ASJ jump test statistics of days identified as having jumps: 1-min sampling frequency.
Notes: Fig. 4 displays the scatter plot for ASJ test statistics for days identified as having jumps using 1-min sampling frequency. See notes of Fig. 3.

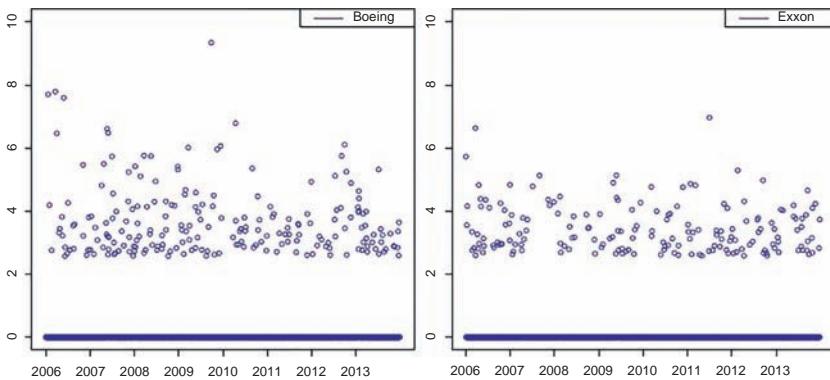


FIG. 5 BNS jump test statistics of days identified as having jumps. Notes: Fig. 5 displays the scatter plot for BNS test statistics for days identified as having jumps using 5-min sampling frequency. See notes of Fig. 3.

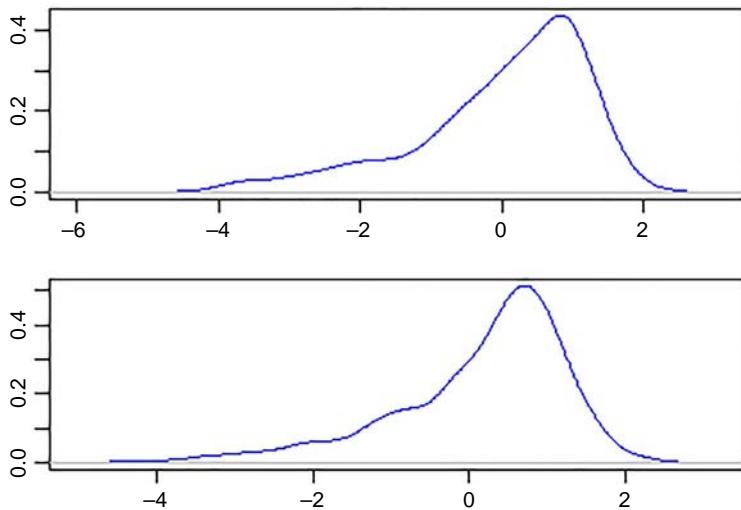


FIG. 6 Kernel density plots for ASJ test statistics. Notes: Fig. 6 displays the kernel density plot of ASJ jump test statistics using 5-min sampling frequency. See notes of Fig. 3.

and LM jump test. Co-jumps are detected in case of each of the following pairwise stock combinations, including Exxon & JPMorgan, Exxon & Microsoft, Exxon & XLE, JPMorgan & Microsoft, Microsoft & XLK and XLE & XLK. The range of percentage of co-jump days in JT co-jump test is from 0.454% to 2.955%, while the range for the co-exceedance rule is from 2.838% to 9.545%. One reason for the larger percentage range in co-exceedance rule could be the fact that the intersection results between two jump tests lead to a large false rejection rate. The percentage of co-jump days in JPMorgan & Microsoft, Microsoft & XLK and XLE & XLK is larger during the financial crisis period

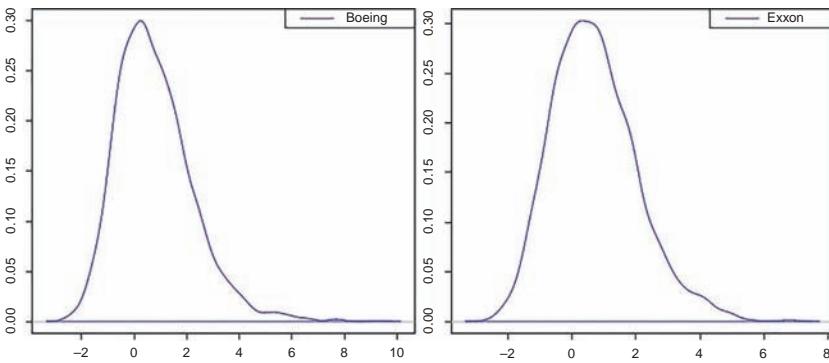


FIG. 7 Kernel density plot of BNS jump test statistics. Notes: Fig. 7 displays the kernel density plot of BNS jump test statistics using 5-min sampling frequency. See notes of Fig. 3.

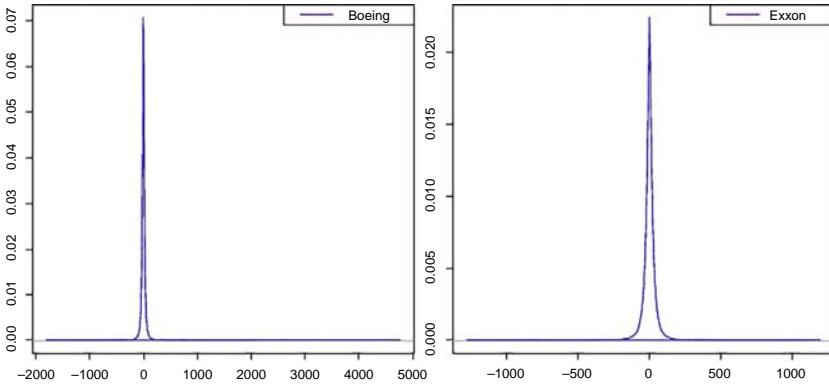


FIG. 8 Kernel density plot of LM jump test statistics. Notes: Fig. 8 displays the kernel density plot of LM jump test statistics using 5-min sampling frequency. See notes of Fig. 3.

than postfinancial crisis period and this result is robust across the different significance levels and types of co-jump tests. In Table 13, we detect co-jumps among the six stocks (Boeing, Exxon, Johnson & Johnson, JPMorgan, Microsoft and Walmart), using the BLT co-jump test as in Bollerslev et al. (2008). As is clear from the table, the percentage of co-jump days as per the BLT test is small, ranging from 0.114% to 0.454%.

We now turn to discuss graphical representation of co-jumps. Figs. 9–11 have only been given for co-jumps between pairs Exxon & JPMorgan, Exxon & Microsoft. Fig. 9 denotes the kernel density plot of JT co-jump test. Overall the distribution of the test statistics appears to be heavily right tailed. Fig. 10 shows the JT test statistics of co-jump days from year 2006 to 2013. It is clear that co-jumps are less densely populated when compared with jump days. When comparing how co-jumps are scattered between financial crisis period and post-financial crisis period, there is no significant difference among Exxon & JPMorgan, Exxon & Microsoft. On the other hand more frequent co-jumps

TABLE 9 Average ratio of jump variation to total variation—ASJ jump test

| Name | Panel A: January 2006–June 2009 | | | | Panel B: July 2009–December 2012 | | | |
|--------------------|---------------------------------|-------|-------|------|----------------------------------|------|-------|------|
| | 1 min | | 5 min | | 1 min | | 5 min | |
| Significance level | 0.05 | 0.10 | 0.05 | 0.10 | 0.05 | 0.10 | 0.05 | 0.10 |
| Boeing | 3.74 | 6.14 | 0.70 | 2.38 | 2.14 | 4.08 | 0.32 | 1.68 |
| Exxon | 2.86 | 4.28 | 0.44 | 1.63 | 1.55 | 3.02 | 0.26 | 1.03 |
| Johnson & Johnson | 5.21 | 8.09 | 0.99 | 2.87 | 3.49 | 5.96 | 0.64 | 2.17 |
| JPMorgan | 2.97 | 5.21 | 0.55 | 2.16 | 1.56 | 2.90 | 0.42 | 1.14 |
| Microsoft | 4.59 | 8.05 | 0.50 | 2.07 | 3.30 | 5.47 | 0.96 | 2.20 |
| Walmart | 3.50 | 6.30 | 0.53 | 2.18 | 2.93 | 5.33 | 0.86 | 1.81 |
| XLE | 2.08 | 3.35 | 0.34 | 1.39 | 1.44 | 2.40 | 0.24 | 1.05 |
| XLK | 10.78 | 16.96 | 1.27 | 3.40 | 4.69 | 7.88 | 0.54 | 1.74 |

Notes: Table 9 gives the average ratio of jump variation to total variation as per the ASJ test using both 5- and 1-min sampling frequencies. Jump ratio is calculated using Eq. (89). Panel A covers the financial crisis period from January 2006–June 2009 and panel B covers the postfinancial crisis period from July 2009–December 2012.

TABLE 10 Average ratio of jump variation to total variation—BNS jump test

| Name | Panel A: January 2006–June 2009 | | Panel B: July 2009–December 2012 | |
|--------------------|---------------------------------|-------|----------------------------------|-------|
| Significance level | 0.05 | 0.10 | 0.05 | 0.10 |
| Boeing | 37.87 | 34.36 | 6.10 | 8.58 |
| Exxon | 32.00 | 28.44 | 6.93 | 7.67 |
| Johnson & Johnson | 42.46 | 39.93 | 9.30 | 11.09 |
| JP Morgan | 36.81 | 33.75 | 5.99 | 8.13 |
| Microsoft | 36.44 | 33.26 | 8.92 | 9.76 |
| Walmart | 36.60 | 33.43 | 7.91 | 8.87 |
| XLE | 33.00 | 30.08 | 4.37 | 5.64 |
| XLK | 44.63 | 41.78 | 12.75 | 14.70 |

Notes: Table 10 shows average ratio of jump variation to total variation as per the BNS test using 5-min frequency. See notes of Table 9.

TABLE 11 Average ratio of jump variation to total variation—LM jump test

| Name | Panel A: January 2006–June 2009 | | Panel B: July 2009–December 2012 | |
|--------------------|---------------------------------|-------|----------------------------------|-------|
| Significance level | 0.05 | 0.10 | 0.05 | 0.10 |
| Boeing | 17.72 | 17.79 | 14.59 | 14.87 |
| Exxon | 14.65 | 14.58 | 12.59 | 12.62 |
| Johnson & Johnson | 21.53 | 21.53 | 19.34 | 19.4 |
| JP Morgan | 17.41 | 17.36 | 13.92 | 13.93 |
| Microsoft | 17.57 | 17.57 | 15.59 | 15.60 |
| Walmart | 17.68 | 17.56 | 15.22 | 15.23 |
| XLE | 13.58 | 13.56 | 10.98 | 11.17 |
| XLK | 25.45 | 25.35 | 21.33 | 21.41 |

Notes: Table 11 shows the average ratio of jump variation to total variation as per the LM test using 5-min frequency. See notes of Table 9.

TABLE 12 Percentage of days identified as having co-jumps

| | Panel A: January 2006–June 2009 | | | | Panel B: July 2009–December 2012 | | | |
|----------------------|---------------------------------|-------|---------------|-------|----------------------------------|-------|---------------|-------|
| | JT Test | | LM & BNS Test | | JT Test | | LM & BNS Test | |
| | 0.05 | 0.10 | 0.05 | 0.10 | 0.05 | 0.10 | 0.05 | 0.10 |
| Exxon & JPMorgan | 1.364 | 1.818 | 3.864 | 7.159 | 1.703 | 1.816 | 2.838 | 5.335 |
| Exxon & Microsoft | 0.795 | 0.909 | 3.295 | 6.818 | 1.022 | 1.476 | 3.519 | 5.675 |
| Exxon & XLE | 1.136 | 1.59 | 3.977 | 7.386 | 1.93 | 2.497 | 3.746 | 8.059 |
| JPMorgan & Microsoft | 1.136 | 1.363 | 3.409 | 5.682 | 1.135 | 1.249 | 2.951 | 5.335 |
| Microsoft & XLK | 2.500 | 2.955 | 6.136 | 9.545 | 1.022 | 1.362 | 4.767 | 7.491 |
| XLE & XLK | 2.045 | 2.386 | 4.773 | 8.295 | 0.454 | 0.568 | 3.065 | 6.129 |

Notes: Table 12 shows the percentage of days identified as having co-jumps. Co-jumps are detected at $\alpha = 0.05$ and $\alpha = 0.1$ significance level. Both JT co-jump test and co-exceedance rule between BNS test and LM test are used to test co-jumps. Panel A covers the financial crisis period from January 2006–June 2009 and panel B covers the postfinancial crisis period from July 2009–December 2012. The test statistics are calculated at 5-min frequency.

TABLE 13 Percentage of days identified as having co-jumps

| | Panel A | | Panel B | |
|---------------------------|---------|-------|---------|-------|
| <i>Significance level</i> | 0.05 | 0.1 | 0.05 | 0.1 |
| BLT Test | 0.227 | 0.341 | 0.114 | 0.454 |

Notes: See notes to Table 12. Table 13 shows the percentage of days identified as having co-jumps from BLT co-jump tests. Co-jumps are detected among six stocks: Boeing, Exxon, Johnson & Johnson, JPMorgan, Microsoft and Walmart.

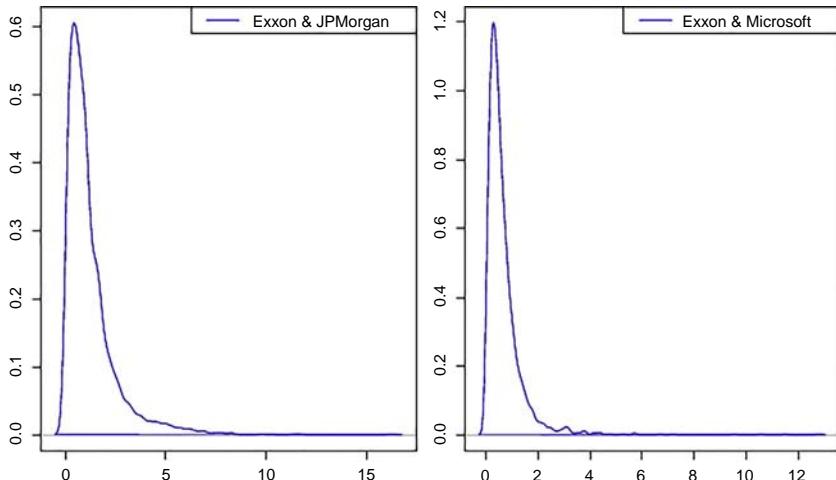


FIG. 9 Kernel density plot of JT co-jump test statistics. Notes: Fig. 9 displays the kernel density plot of JT co-jump test using 5-min sampling frequency. The co-jumps are tested for the pairs Exxon & JPMorgan and Exxon & Microsoft for the sample period January 2006–December 2013.

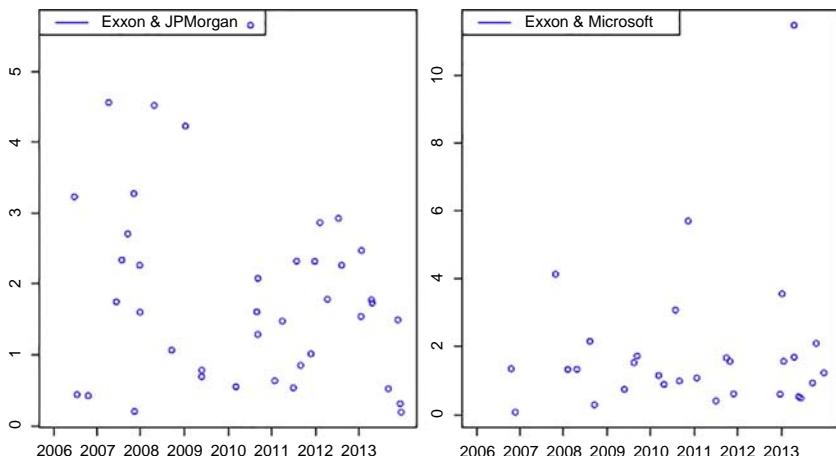


FIG. 10 JT co-jump test statistics of days identified as having co-jumps. Notes: Fig. 10 displays the co-jump days test statistics of JT test for the sample period January 2006–December 2013 using sampling frequency of 5 min.

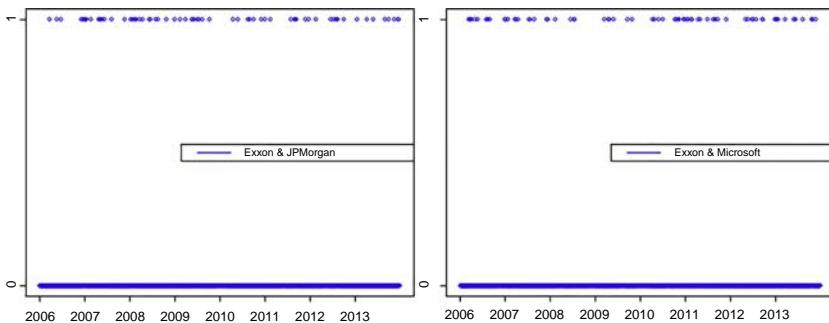


FIG. 11 LM & BNS test statistics for days having co-jumps. Notes: Fig. 11 displays the co-jump days identified from co-exceedance rule between LM jumps test and BNS jump test for the pairs Exxon & JPMorgan and Exxon & Microsoft. 5-min sampling frequency is considered for sample period January 2006–December 2013.

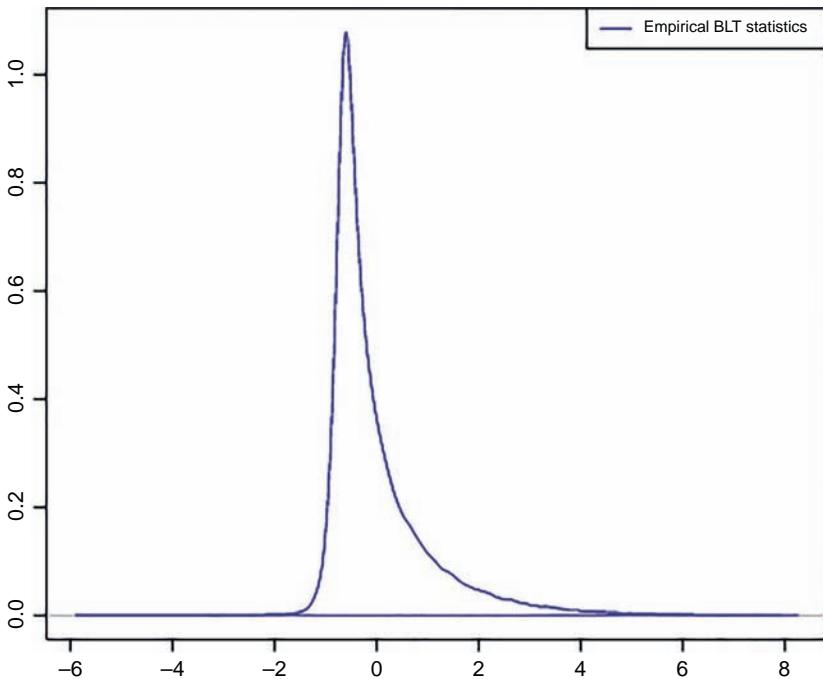


FIG. 12 Kernel density plot of empirical observed BLT statistics. Notes: Fig. 12 displays the kernel density plot of the empirical observed BLT co-jump test statistics using 5-min sampling frequency for the sample period January 2006–December 2013.

are visible during the financial crisis period in Microsoft & XLK and XLE & XLK. [Fig. 11](#) shows the days which have co-jumps as per the co-exceedance rule. The results show there is not much significant difference on how co-jumps are distributed between financial crisis and postfinancial crisis period.

Finally, [Figs. 12 and 13](#) show the empirical findings from BLT co-jump tests. [Fig. 12](#) denotes the kernel density plot of empirical BLT test statistics.

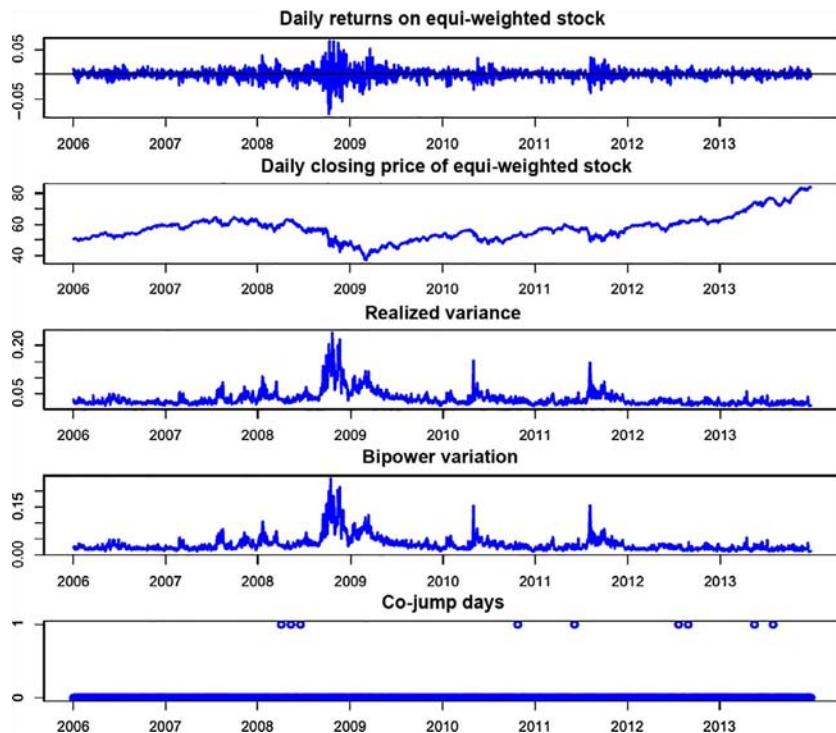


FIG. 13 Daily return, daily closing price, realized variance, bipower variation, and co-jump days for equi-weighted stock index. Notes: Fig. 13 displays the daily return, daily closing price, realized variance, bipower variation and co-jump days for equi-weighted stock index. The co-jump days in the last panel are detected through BLT co-jump tests at $\alpha = 0.1$ significance level from January 2006 to December 2013. The equi-weighted stock index is composed of six stocks (Boeing, Exxon, Johnson & Johnson, JPMorgan, Microsoft, and Walmart) with equal weights.

The distribution of the test statistics is evidently positively skewed. Fig. 13 shows the daily return, daily closing price, realized variance, BPV, and co-jump days for equi-weighted stock index. In Bollerslev et al. (2008), the authors show that detection of co-jumps among multiple stocks is equivalent to detecting co-jumps in an equi-weighted index composed by the same underlying stocks. Here we test co-jumps among six stocks, including Boeing, Exxon, Johnson & Johnson, JPMorgan, Microsoft and Walmart. The last panel of Fig. 13 shows the number of co-jump days at $\alpha = 10\%$ significance level. There are only 9 co-jump days among six stocks from year 2006 to 2013.

7 Conclusion

In this chapter, we review some of the most recent literature on integrated volatility measures, jump and co-jump tests. We then select a small subset of these

measures and tests to conduct an empirical investigation with intra-day TAQ data of six individuals stocks and two ETFs. This study helps to reveal how the general volatility movement, jump and co-jump activity among the stocks vary across different types of tests and sampling frequencies.

We find that the occurrence of jumps is more frequent during and before the financial crisis period, i.e., January 2006–June 2009 compared to the postfinancial crisis period, i.e., July 2009–December 2013. All individual stocks apart from the ETFs reach their peak volatility in the fourth quarter of 2008. Overall, the incidence of co-jumps is lesser compared to jumps over the entire sample period, i.e., January 2006–December 2013. Additionally there is not much significant difference in terms of distribution of co-jumps between financial crisis and postfinancial crisis period.

Appendix. R code

Please find below R (statistical software) codes for the tests and measures which have been used in the empirical section of this paper. This includes the volatility measures *RV*, *BPV*, *TPV*, *MinRV*, *MedRV*, *TRV*; ASJ, BNS jump tests; and BLT and JT co-jump tests. The format of input data can be given as follows: rows signify the trading days and the columns signify the intra-day intervals.

```
#-----#
      #### Integrated Volatility Measures ####
#-----#
##-----##
## Data preparation 1 ##
##-----##
data <- read.csv(file="asset1_1min_data.csv", header=FALSE,
sep=",")
mat <- data.matrix(data)
fin_data <- t(mat)
days <- nrow(data)
freq <- 79
data_5 <- matrix(0, freq, days)
for (j in 1:days){for(i in 1:freq){data_5[i,j]
<- fin_data[(i-1)*5+1,j]}}
##-----##
## Realized Volatility - RV ##
##-----##
dif_data <- diff(data_5)
RV <- colSums((dif_data)^2)
```

```

## -----
## Bipower Variation - BPV ##
## -----
dif1 <- dif_data[1:freq-2,1:days]
dif2 <- dif_data[2:freq-1,1:days]
BPV <- (sqrt(2)/sqrt(pi))*colSums(abs(dif1)*abs(dif2))

## -----
## Tripower variation - TPV ##
## -----
dif11 <- dif_data[1:freq-3, 1:days]
dif22 <- dif_data[2:freq-2, 1:days]
dif33 <- dif_data[3:freq-1, 1:days]
cons <- (((2^(1/3))*gamma(5/6))/gamma(1/2))^-3
TPV <- cons*colSums((abs(dif11)^(2/3))*(abs(dif22)
^(2/3))*(abs(dif33)^(2/3)))

## -----
## MinRV ##
## -----
minvec <- matrix(0, freq-2, days)
for (j in 1:days){for (i in 1:freq-2){minvec[i,j]
<- (min( abs(dif1[i,j]), abs(dif2[i,j]) ))^2}}
MinRV <- (pi/(pi-2))*(freq/freq-1)*colSums(minvec)

## -----
## MedRV ##
## -----
medvec <- matrix(0, freq-3, days)
for (j in 1:days){for (i in 1:freq-3){medvec[i,j]
<- (median(c(abs(dif11[i,j]),abs(dif22[i,j]),abs(dif33[i,j]))))^2}}
MedRV <- (pi/(6 - 4*sqrt(3) + pi))*(freq/freq-2)*colSums(medvec)

## -----
## Truncated Realized Volatility - TRV ##
## -----
delta <- 1/freq
omega <- 0.47
alpha <- matrix(0, freq-1, days)
for (j in 1: days){
  for (i in 1: freq-1){if (abs(dif_data[i,j]) <= sqrt(delta))
    {alpha[i,j] <- abs((dif_data[i,j]))^2} else {alpha[i,j] <- 0}}
  alph_fin <- 5*sqrt(colSums(alpha))
  trun <- matrix(0, freq-1, days)
  for (j in 1: days){for (i in 1: freq-1){
    if (abs(dif_data[i,j]) <= alph_fin[j]*delta^omega)
      {trun[i,j] <- abs((dif_data[i,j]))^2} else {trun[i,j] <- 0}}}
  TRV <- colSums(trun)
}

```

```

##-----##  

##### Jump Tests #####  

##-----##  

##-----##  

## ASJ Jump Test ##  

##-----##  

BPD <- colSums((abs(dif_data))^4)  

kfreq = (freq+1)/2  

data_10 <- matrix(0, kfreq, days)  

for (j in 1: days){for (i in 1:kfreq){data_10[i,j]  

<- data_5[(i-1)*2+1,j]}}  

BPK <- colSums((abs(diff(data_10)))^4)  

SPK <- BPK/BPD  

trun_4 <- matrix(0,freq-1,days)  

for (j in 1: days){for (i in 1:freq-1){  

  if (abs(dif_data[i,j]) <= alph_fin[j]*delta^omega)  

{trun_4[i,j] <- abs((dif_data[i,j]))^4}else {trun_4[i,j] <- 0}}}  

mp <- pi^(-0.5)*4*gamma(5/2)  

AP <- (delta^(-1)/mp)*colSums(trun_4)  

trun_8 <- matrix(0, freq-1, days)  

for (j in 1: days){for (i in 1: freq-1){  

  if (abs(dif_data[i,j]) <= alph_fin[j]*delta^omega)  

{trun_8[i,j] <- abs((dif_data[i,j]))^8} else  

{trun_8[i,j] <- 0}}}  

mp_8 <- pi^(-0.5)*16*gamma(9/2)  

AP_8 <- (delta^(-3)/mp_8)*colSums(trun_8)  

Var <- (delta* AP_8*160)/(3*AP^2)  

ASJ <- (2 - SPK)/sqrt(Var)  

##-----##  

## Data preparation 2 ##  

##-----##  

data_asset1 <- read.table("asset1_1min_data.csv",  

header=FALSE,sep=",");  

data_asset2 <- read.table("asset2_1min_data.csv",  

header=FALSE, sep=",");  

data_asset3 <- read.table("asset3_1min_data.csv",  

header=FALSE,sep=",");  

col_diff_asset1 <- apply(data_asset1, 1, diff);  

col_diff_asset2 <- apply(data_asset2, 1, diff);  

col_diff_asset3 <- apply(data_asset3, 1, diff);  

vec_asset1=as.vector(col_diff_asset1);  

vec_asset2=as.vector(col_diff_asset2);  

vec_asset3=as.vector(col_diff_asset3);  

vec_asset1_5min=colSums(matrix(vec_asset1, nrow=5));  

vec_asset2_5min=colSums(matrix(vec_asset2, nrow=5));  

vec_asset3_5min=colSums(matrix(vec_asset3, nrow=5));

```

```

matrix_5min_all=cbind(vec_asset1_5min, vec_asset2_5min,
vec_asset3_5min);

##-----##
## BNS Jump Test ##
##-----##
simu_n <- days;
N=ncol(matrix_5min_all);
freq=nrow(matrix_5min_all)/simu_n;
RVdaily=matrix(0,nrow=simu_n,ncol=N);
matrix_5min_all_sqr=matrix_5min_all*matrix_5min_all;
for (j in 1:N){for (i in 1:simu_n){RVdaily[i,j]=
sum(matrix_5min_all_sqr[((i-1)*freq+1):(i*freq),j])}}
BV=matrix(0,nrow=simu_n,ncol=N);
mul=sqrt(2/pi);
for (k in 1:N){ for (i in 1:simu_n){for (j in 2:freq)
(BV[i,k]=BV[i,k]+abs(matrix_5min_all[((i-1)*freq+j),
k])*abs(matrix_5min_all[((i-1)*freq+j-1),k]))}
BV[i,k]=(mul)^(-2)*(freq/(freq-1))*BV[i,k]}}
QP=matrix(0,nrow=simu_n,ncol=N);
for (k in 1:N){for (i in 1:simu_n){for (j in 4: freq){
QP[i,k]=QP[i,k]+(abs(matrix_5min_all[((i-1)*freq+j),k]))*
(abs(matrix_5min_all[((i-1)*freq+j-1),k]))*
(abs(matrix_5min_all[((i-1)*freq+j-2),k]))*
(abs(matrix_5min_all[((i-1)*freq+j-3),k]));}
QP[i,k]=freq*QP[i,k]}}
Vqq=2;
Vbb=(pi/2)^2+pi-3;
Zqp=matrix(0,nrow=simu_n,ncol=N);
for (k in 1:N){for (i in 1:simu_n){
Zqp[i,k]=-(((BV[i,k]/RVdaily[i,k])-1)*(freq^0.5))/(
(sqrt((Vbb-Vqq)*max(1,QP[i,k]/(BV[i,k]^2*mul^4)))) )}}
detect_BNS=matrix(0,nrow=simu_n,ncol=N);
for (k in 1:N){for (j in 1:simu_n)
{detect_BNS[j,k]=ifelse(abs(Zqp[j,k])>1.96,1,0)}}
```

##-----##
Co-Jump Tests ##
##-----##

```

##-----##
## BLT Co-Jump Test ##
##-----##
library(MASS)
Input1=matrix_5min_all
cov_cof=cov(matrix_5min_all)
N=ncol(matrix_5min_all)
simu_n<-days;
mu=rep(0,N);

```

```

freq=nrow(matrix_5min_all)/simu_n;
set.seed(123)
bpath=mvrnorm(n = simu_n*freq, mu, cov_cof, empirical = FALSE)
dt=1/freq
sigma<-10
bdw=sqrt(dt)*bpath*sigma
bz=matrix(0,nrow=simu_n*freq,ncol=N-1)
for(i in 1:(N-1)){for(l in (i+1):N){bz[,i]=bz[,i]+bdw[,i]*bdw[,l]}}
bmcp=rowSums(bz)
bmcptj=(2/((N-1)*N))*bmcp
bmcp_bar=matrix(0,nrow=simu_n,ncol=1)
for (i in 1:simu_n){bmcp_bar[i]=mean(bmcptj[((i-1)*freq+1):
(i*freq)])}
mcp_bar1=matrix(0,nrow=simu_n*freq,1)
for (i in 1:simu_n){mcp_bar1[((i-1)*freq+1):
(i*freq),]=bmcp_bar[i]}
mcp_sqr=(bmcptj-mcp_bar1)*(bmcptj-mcp_bar1)/(freq-1)
sd_mcpt=matrix(0,nrow=simu_n,ncol=1)
for (i in 1:simu_n){sd_mcpt[i]=sqrt(sum(mcp_sqr[((i-1)*freq+1):
(i*freq)]))}
mcp_sqr1=matrix(0,nrow=freq*simu_n,ncol=1)
for (i in 1:simu_n){mcp_sqr1[((i-1)*freq+1):
(i*freq)]=sd_mcpt[i]}
zmcptj=(bmcptj-mcp_bar1)/mcp_sqr1
sortnul=sort(zmcptj, decreasing = TRUE)
rightside=matrix(0,nrow=2,1)
rightside[1]=sortnul[round(freq*simu_n*0.05, digits = 0)]
rightside[2]=sortnul[round(freq*simu_n*0.025, digits = 0)]
sortnull=sort(zmcptj, decreasing = FALSE)
leftside=matrix(0,nrow=2,1)
leftside[1]=sortnull1[round(freq*simu_n*0.05, digits = 0)]
leftside[2]=sortnull1[round(freq*simu_n*0.025, digits = 0)]
z_obs=matrix(0,nrow=simu_n*freq,ncol=N-1)
for (i in 1:(N-1)){for (l in (i+1):N){z_obs[,i]=bz[,i]+Input1
[,i]*Input1[,l]}}
mcp_obs=rowSums(z_obs)
mcptj_obs=(2/((N-1)*N))*mcp_obs
mcp_obsbar=matrix(0,nrow=simu_n,ncol=1)
for (i in 1:simu_n){mcp_obsbar[i]=mean(mcptj_obs[((i-1)*freq+1):
(i*freq)])}
mcp_obsbar1=matrix(0,nrow=freq*simu_n,ncol=1)
for (i in 1:simu_n){mcp_obsbar1[((i-1)*freq+1):
(i*freq),]=mcp_obsbar[i]}
mcp_obsqr=(mcptj_obs-mcp_obsbar1)*(mcptj_obs-mcp_obsbar1)/
(freq-1)
Sd_obsmcpt=matrix(0,nrow=simu_n,ncol=1)

```

```

for (i in 1:simu_n){Sd_obsmpct[i]=sqrt(sum(mcp_obsqr
[((i-1)*freq+1):(i*freq])))}
mcp_obsqr1=matrix(0,nrow=freq*simu_n,ncol=1)
for (i in 1:simu_n){mcp_obsqr1[((i-1)*freq+1):
(i*freq)]=Sd_obsmpct[i]}
zmcptj_obs=(mcptj_obs-mcp_obsbar1)/mcp_obsqr1
detectr=matrix(0,nrow=freq*simu_n,ncol=4)
for (i in 1:4){ for (j in 1:freq*simu_n){
detectr[j,i]=ifelse(zmcptj_obs[j]>rightside[i], 1, 0))}
detectl=matrix(0,nrow=freq*simu_n,ncol=4)
for (i in 1:4){ for (j in 1:freq*simu_n){
detectl[j,i]=ifelse(zmcptj_obs[j]<leftside[i], 1, 0))}

##-----##
## JT Co-Jump Test ##
##-----##
library(MASS)
combos=combn(1:ncol(matrix_5min_all), 2, FUN = NULL,
simplify = TRUE);
n1=ncol(combos);
V=matrix(0,nrow=simu_n,ncol=n1);
Input1=matrix_5min_all;
freq=nrow(matrix_5min_all)/simu_n;
for (m in 1:n1){for (j in 1:simu_n){
for (i in 1: freq){V[j,m]=V[j,m]+(Input1[((j-1)*
freq)+i,combos[1,m]]*Input1[((j-1)*freq+i,combos[2,m]])^2)}};
V_g1=matrix(0,nrow=simu_n,ncol=n1);
for (m in 1:n1){for (j in 1:simu_n){for (i in 1: freq)
{V_g1[j,m]=V_g1[j,m]+(Input1[((j-1)*freq+i,combos[1,m]])^4)}};
V_g2=matrix(0,nrow=simu_n,ncol=n1);
for (m in 1:n1){for (j in 1:simu_n){for (i in 1: freq)
{V_g2[j,m]=V_g2[j,m]+(Input1[((j-1)*freq+i,combos[2,m]])^4)}};
phi_d=V/sqrt(V_g1*V_g2);
delta=1/freq;
A_hatn=matrix(0,nrow=simu_n,ncol=n1);
for (m in 1:n1){for (j in 1:simu_n){for (i in ((j-1)*freq+1):
((j-1)*freq+freq-3)){A_hatn[j,m]=A_hatn[j,m]+abs(Input1
[i,combos[1,m]]
*Input1[i+1,combos[1,m]]*Input1[i+2,combos[2,m]]
*Input1[i+3,combos[2,m]])
+(1/8)*abs((Input1[i,combos[1,m]]+Input1
[i,combos[2,m]])*(Input1[i+1,combos[1,m]]
+Input1[i+1,combos[2,m]])*(Input1[i+2,combos[1,m]]
+Input1[i+2,combos[2,m]])
*(Input1[i+3,combos[1,m]]+Input1[i+3,combos[2,m]]));
+(1/8)*abs((Input1[i,combos[1,m]]
-Input1[i,combos[2,m]])
*(Input1[i+1,combos[1,m]]-Input1[i+1,combos[2,m]]))
}}
```

```

*(Input1[i+2, combos[1,m]]-Input1[i+2, combos[2,m]]))
*(Input1[i+3, combos[1,m]]-Input1[i+3, combos[2,m]]));
-(1/4)*abs((Input1[i, combos[1,m]]
+Input1[i, combos[2,m]]])
*(Input1[i+1, combos[1,m]]+Input1[i+1, combos[2,m]])
*(Input1[i+2, combos[1,m]]
-Input1[i+2, combos[2,m]])*(Input1[i+3, combos[1,m]]
-Input1[i+3, combos[2,m]]));
})
}
A_hatn=A_hatn*(pi)^2/(4*delta);
BV=matrix(0,nrow=simu_n,ncol=ncol(matrix_5min_all));
mu1=sqrt(2/pi);
for (k in 1:5){for (i in 1:simu_n){ for (j in 2:freq)
{BV[i,k]=BV[i,k]+abs(Input1[((i-1)*freq+j),k])*abs(Input1[((i-1)
*freq+j-1),k])}}}
BV=(mu1)^(-2)*(freq/(freq-1))*BV
trunc=1*sqrt(BV)*delta^(0.49);
mu=c(0,0);
sigma=matrix(c(1, 0, 0,1), nrow=2, ncol=2);
Nn=20
kn=1/sqrt(delta)
Z_alphaad_10=matrix(0,nrow=simu_n,ncol=n1);
Z_alphaad_5=matrix(0,nrow=simu_n,ncol=n1);
for (m in 1:n1){
  for (kk in 1:simu_n){
    Dhatn=matrix(0,nrow=Nn,ncol=1)
    chatp1=matrix(0,nrow=freq-2*round(kn)-1,ncol=1)
    for (i in ((kk-1)*freq+1+round(kn)):((kk-1)*freq+freq-round(kn)-1)){
      for (j in (i+1):(i+round(kn)+1)){if (abs
      (Input1[j, combos[1,m]])<=trunc[kk, combos[1,m]]){
        chatp1[i-round(kn)-(kk-1)*freq,]=chatp1
        [i-round(kn)-(kk-1)*freq,]
        +Input1[j, combos[1,m]]*Input1[j, combos[1,m]]}
      else{chatp1[(i-round(kn)-(kk-1)*freq),]=chatp1
        [i-round(kn)-(kk-1)*freq,]}
      chatp1[i-round(kn)-(kk-1)*freq,]=(1/(kn*delta))
      *chatp1[i-round(kn)-(kk-1)*freq,]}
      chatm1=matrix(0,nrow=freq-2*round(kn)-1,ncol=1)
      for (i in ((kk-1)*freq+1+round(kn)):((kk-1)
      *freq+freq-round(kn)-1)){
        for (j in (i-round(kn)):(i-1)){if (abs(Input1
        [j, combos[1,m]])<=trunc[kk, combos[1,m]]){
          chatm1[i-round(kn)-(kk-1)*freq,]=chatm1
          [i-round(kn)-(kk-1)*freq,]
          +Input1[j, combos[1,m]]*Input1[j, combos[1,m]]}
        else{chatm1[(i-round(kn)-(kk-1)*freq),]=chatm1
          [i-round(kn)-(kk-1)*freq,]}}}
    }
  }
}
```

```

chatm1[i-round(kn)-(kk-1)*freq,]=(1/(kn*delta))
*chatm1[i-round(kn)-(kk-1)*freq])
chatp2=matrix(0,nrow=freq-2*round(kn)-1,ncol=1)
for (i in ((kk-1)*freq+1+round(kn)):((kk-1)
*freq+freq-round(kn)-1)){
  for (j in (i+1):(i+round(kn)+1)){if (abs(Input1
[j, combos[2,m]])<=trunc[kk, combos[2,m]]){
    chatp2[i-round(kn)-(kk-1)*freq,]=chatp1
[i-round(kn)-(kk-1)*freq,]
+Input1[j, combos[2,m]]*Input1[j, combos[2,m]]}
  else{chatp2[(i-round(kn)-(kk-1)*freq,)] =chatp2
[i-round(kn)-(kk-1)*freq,]}}
  chatp2[i-round(kn)-(kk-1)*freq,]=(1/(kn*delta))
  *chatp2[i-round(kn)-(kk-1)*freq,])
  chatm2=matrix(0,nrow=freq-2*round(kn)-1,ncol=1)
  for (i in ((kk-1)*freq+1+round(kn)):((kk-1)
*freq+freq-round(kn)-1)){
    for (j in (i-round(kn)):(i-1)){if (abs(Input1[j,
combos[2,m]])<=trunc[kk, combos[2,m]]){
      chatm2[i-round(kn)-(kk-1)*freq,]=chatm2
[i-round(kn)-(kk-1)*freq,]
+Input1[j, combos[2,m]]*Input1[j, combos[2,m]]}
    else{chatm2[(i-round(kn)-(kk-1)*freq,)] =chatm2
[i-round(kn)-(kk-1)*freq,]}}
    chatm2[i-round(kn)-(kk-1)*freq,]=(1/(kn*delta))
    *chatm2[i-round(kn)-(kk-1)*freq,])
    for (s in 1:Nn){set.seed(s);
    U=matrix(0,nrow=freq-2*round(kn)-1,ncol=2);
    for (i in 1:(freq-2*round(kn)-1)){U[i,]=mvtnorm
(n = 1, mu, sigma)};
    Up=matrix(0,nrow=freq-2*round(kn)-1,ncol=2)
    for (i in 1:(freq-2*round(kn)-1)){Up[i,]=mvtnorm
(n = 1, mu, sigma)}
    K=runif(freq-2*round(kn)-1);
    R1=sqrt(K)*sqrt(chatm1)*U[,1]+sqrt(1-K)
    *sqrt(chatp1)*Up[,1];
    R2=sqrt(K)*sqrt(chatm2)*U[,1]+sqrt(1-K)
    *sqrt(chatp2)*Up[,1];
    for (i in ((kk-1)*freq+1+round(kn)):((kk-1)
*freq+freq-round(kn)-1)){
      if (abs(Input1[i, combos[1,m]])>trunc
[kk, combos[1,m]]
      && abs(Input1[i, combos[2,m]])>trunc
[kk, combos[2,m]]){
        Dhatn[s]=Dhatn[s]+(Input1[i, combos[1,m]]
*R2[i-round(kn)-(kk-1)*freq])^2
      }
    }
  }
}

```

```

+(Input1[i, combos[2,m]]*R1[i-round(kn)-(kk-1)
 *freq])^2}else(Dhatn[s]=Dhatn[s]))}
Dhatns=sort(Dhatn, decreasing = TRUE);
Z_alpha10_10[kk,m]=Dhatns[Nn*0.1];
Z_alpha10_5[kk,m]=Dhatns[round(Nn*0.05)];
}
cnd_10=sqrt(delta)*(Z_alpha10_10+A_hatn)/sqrt(V_g1*V_g2);
cnd_5=sqrt(delta)*(Z_alpha10_5+A_hatn)/sqrt(V_g1*V_g2);
##-----##

```

References

- Aït-Sahalia, Y., 2002. Telling from discrete data whether the underlying continuous-time model is a diffusion. *J. Financ.* 57, 2075–2112.
- Aït-Sahalia, Y., Jacod, J., et al., 2009. Testing for jumps in a discretely observed process. *Ann. Stat.* 37, 184–222.
- Andersen, T.G., Bollerslev, T., Diebold, F.X., Labys, P., 2001. The distribution of realized exchange rate volatility. *J. Am. Stat. Assoc.* 96, 42–55.
- Andersen, T.G., Benzoni, L., Lund, J., 2002. An empirical investigation of continuous-time equity return models. *J. Financ.* 57, 1239–1284.
- Andersen, T.G., Bollerslev, T., Diebold, F.X., 2003. Some like it smooth, and some like it rough: Untangling continuous and jump components in measuring, modeling, and forecasting asset return volatility. In: CFS Working Paper Series 2003/35. Center for Financial Studies (CFS).
- Andersen, T.G., Bollerslev, T., Diebold, F.X., 2007. Roughing it up: including jump components in the measurement, modeling, and forecasting of return volatility. *Rev. Econ. Stat.* 89, 701–720.
- Andersen, T.G., Bollerslev, T., Frederiksen, P., Ørregaard Nielsen, M., 2010. Continuous-time models, realized volatilities, and testable distributional implications for daily stock returns. *J. Appl. Econom.* 25, 233–261.
- Andersen, T.G., Dobrev, D., Schaumburg, E., 2012. Jump-robust volatility estimation using nearest neighbor truncation. *J. Econom.* 169, 75–93.
- Bandi, F.M., Reno, R., 2016. Price and volatility co-jumps. *J. Financ. Econ.* 119, 107–146.
- Barndorff-Nielsen, O.E., Shephard, N., 2004. Power and bipower variation with stochastic volatility and jumps. *J. Financ. Econom.* 2, 1–37.
- Barndorff-Nielsen, O.E., Shephard, N., 2006. Econometrics of testing for jumps in financial economics using bipower variation. *J. Financ. Econom.* 4, 1–30.
- Barndorff-Nielsen, O.E., Shephard, N., et al., 2003. Realized power variation and stochastic volatility models. *Bernoulli* 9, 243–265.
- Barndorff-Nielsen, O.E., Graversen, S.E., Jacod, J., Podolskij, M., Shephard, N., 2006. A central limit theorem for realised power and bipower variations of continuous semimartingales. In: From Stochastic Calculus to Mathematical Finance. Springer, pp. 33–68.
- Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2008. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* 76, 1481–1536.
- Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A., Shephard, N., 2011. Subsampling realised kernels. *J. Econom.* 160, 204–219.

- Bibinger, M., Winkelmann, L., 2015. Econometrics of co-jumps in high-frequency data with noise. *J. Economet.* 184, 361–378.
- Bollerslev, T., Law, T.H., Tauchen, G., 2008. Risk, jumps, and diversification. *J. Economet.* 144, 234–256.
- Boswijk, H.P., Laeven, R.J.A., Yang, X., 2018. Testing for self-excitation in jumps. *J. Economet.* 203, 256–266.
- Boudt, K., Croux, C., Laurent, S., 2011. Robust estimation of intraweek periodicity in volatility and jump detection. *J. Empir. Financ.* 18, 353–367.
- Brownlees, C.T., Nualart, E., Sun, Y., 2016. A truncated two-scales realized volatility estimator.
- Caporin, M., Kolokolov, A., Renò, R., 2017. Systemic co-jumps. *J. Financ. Econ.* 126, 563–591.
- Carr, P., Wu, L., 2003. What type of process underlies options? A simple robust test. *J. Financ.* 58, 2581–2610.
- Chernov, M., Gallant, A.R., Ghysels, E., Tauchen, G., 2003. Alternative models for stock price dynamics. *J. Economet.* 116, 225–257.
- Corradi, V., Distaso, W., Swanson, N.R., 2011. Predictive inference for integrated volatility. *J. Am. Stat. Assoc.* 106, 1496–1512.
- Corradi, V., Silvapulle, M., Swanson, N., 2014. Consistent pretesting for jumps. https://papers.ssrn.com/sol3/papers.cfmabst_id=2446970.
- Corradi, V., Silvapulle, M.J., Swanson, N.R., 2018. Testing for jumps and jump intensity path dependence. *J. Economet.* 204, 248–267.
- Corsi, F., Pirino, D., Reno, R., 2010. Threshold bipower variation and the impact of jumps on volatility forecasting. *J. Economet.* 159, 276–288.
- Dungey, M., Henry, Ó.T., Hvozdyk, L., 2011. The Impact of Thin Trading and Jumps on Realized Hedge Ratios. University of Cambridge. (manuscript, CFAP).
- Dungey, M.H., Erdemlioglu, D., Matei, M., Yang, X., 2016. Financial flights, stock market linkages and jump excitation. https://www.cb.cityu.edu.hk/ef/doc/2016%20Sofie/Papers/JFQTHKFM_JEX_rv.pdf.
- Duong, D., Swanson, N.R., 2011. Volatility in discrete and continuous-time models: a survey with new evidence on large and small jumps. In: Drukker, D.M. (Ed.), *Missing Data Methods: Time-Series Methods and Applications*, pp. 179–233.
- Eraker, B., Johannes, M., Polson, N., 2003. The impact of jumps in volatility and returns. *J. Financ.* 58, 1269–1300.
- Fan, J., Wang, Y., 2007. Multi-scale jump and volatility analysis for high-frequency financial data. *J. Am. Stat. Assoc.* 102, 1349–1362.
- Gilder, D., Shackleton, M.B., Taylor, S.J., 2014. Cojumps in stock prices: empirical evidence. *J. Bank. Financ.* 40, 443–459.
- Gnabo, J.-Y., Hvozdyk, L., Lahaye, J., 2014. System-wide tail comovements: a bootstrap test for cojump identification on the S&P 500, US bonds and currencies. *J. Int. Money Financ.* 48, 147–174.
- Hausman, J., 1978. Specification tests in econometrics. *Econometrica* 46, 1251–1271.
- Hautsch, N., Podolskij, M., 2013. Preaveraging-based estimation of quadratic variation in the presence of noise and jumps: theory, implementation, and empirical evidence. *J. Bus. Econ. Stat.* 31, 165–183.
- Huang, X., Tauchen, G., 2005. The relative contribution of jumps to total price variance. *J. Financ. Econom.* 3, 456–499.
- Jacod, J., Todorov, V., 2009. Testing for common arrivals of jumps for discretely observed multidimensional processes. *Ann. Stat.* 37, 1792–1838.

- Jacod, J., Rosenbaum, M., et al., 2013. Quarticity and other functionals of volatility: efficient estimation. *Ann. Stat.* 41, 1462–1484.
- Jacod, J., Todorov, V., et al., 2014. Efficient estimation of integrated volatility in presence of infinite variation jumps. *Ann. Stat.* 42, 1029–1069.
- Jacod, J., Li, Y., Zheng, X., 2017. Estimating the integrated volatility with tick observations. <https://www.sciencedirect.com/science/article/pii/S0304407618301714>.
- Jiang, G.J., Oomen, R.C.A., 2008. Testing for jumps when asset prices are observed with noise-a “swap variance” approach. *J. Economet.* 144, 352–370.
- Jing, B.-Y., Liu, Z., Kong, X.-B., 2014. On the estimation of integrated volatility with jumps and microstructure noise. *J. Bus. Econ. Stat.* 32, 457–467.
- Johannes, M., 2004. The statistical and economic role of jumps in continuous-time interest rate models. *J. Financ.* 59, 227–260.
- Kalnina, I., Linton, O., 2008. Estimating quadratic variation consistently in the presence of endogenous and diurnal measurement error. *J. Economet.* 147, 47–59.
- Lahaye, J., Laurent, S., Neely, C.J., 2011. Jumps, cojumps and macro announcements. *J. Appl. Economet.* 26, 893–921.
- Lee, S.S., Mykland, P.A., 2007. Jumps in financial markets: a new nonparametric test and jump dynamics. *Rev. Financ. Stud.* 21, 2535–2563.
- Lee, T., Loretan, M., Ploberger, W., 2013. Rate-optimal tests for jumps in diffusion processes. *Stat. Pap.* 54, 1009–1041.
- Mancini, C., 2001. Disentangling the jumps of the diffusion in a geometric jumping Brownian motion. *Giornale dell'Istituto Italiano degli Attuari* 64, 44.
- Mancini, C., 2009. Non-parametric threshold estimation for models with stochastic diffusion coefficient and jumps. *Scand. J. Stat.* 36, 270–296.
- Mancini, C., Gobbi, F., 2012. Identifying the Brownian covariation from the co-jumps given discrete observations. *Economet. Theor.* 28, 249–273.
- Mukherjee, A., Swanson, N., 2018. New direction for Volatility Confidence Interval Prediction: Evidence From an Experimental and Empirical Study (Working Paper). Rutgers University.
- Mykland, P.A., Zhang, L., 2009. Inference for continuous semimartingales observed at high frequency. *Econometrica* 77, 1403–1445.
- Pan, J., 2002. The jump-risk premia implicit in options: evidence from an integrated time-series study. *J. Financ. Econ.* 63, 3–50.
- Peng, W., Swanson, N.R., 2018. Co-Jumps, Co-Jump Tests and Volatility Prediction in the S&P 500 Market (Working Paper).
- Podolskij, M., Ziggel, D., 2010. New tests for jumps in semimartingale models. *Stat. Infer. Stoch. Process* 13, 15–41.
- Podolskij, M., Vetter, M., et al., 2009. Estimation of volatility functionals in the simultaneous presence of microstructure noise and jumps. *Bernoulli* 15, 634–658.
- Todorov, V., Tauchen, G., 2012. The realized Laplace transform of volatility. *Econometrica* 80, 1105–1127.
- Zhang, L., Mykland, P.A., Aït-Sahalia, Y., 2005. A tale of two time scales: determining integrated volatility with noisy high-frequency data. *J. Am. Stat. Assoc.* 100, 1394–1411.
- Zhang, L., et al., 2006. Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach. *Bernoulli* 12, 1019–1043.

Chapter 2

Real time monitoring of asset markets: Bubbles and crises

Peter C.B. Phillips^{a,b,*} and Shuping Shi^c

^aDepartment of Economics, Yale University, New Haven, CT, United States

^bDepartment of Economics, University of Auckland, Auckland, New Zealand

^cDepartment of Economics, Macquarie University, North Ryde, NSW, Australia

*Corresponding author: e-mail: peter.phillips@yale.edu

Abstract

While each financial crisis has its own characteristics there is now widespread recognition that crises arising from sources such as financial speculation and excessive credit creation do inflict harm on the real economy. Detecting speculative market conditions and ballooning credit risk in real time is therefore of prime importance in the complex exercises of market surveillance, risk management, and policy action. This chapter provides an R implementation of the popular real-time monitoring strategy proposed by Phillips et al. (2015a,b), along with a new bootstrap procedure designed to mitigate the potential impact of heteroskedasticity and to effect family-wise size control in recursive testing algorithms. This methodology has been shown effective for bubble and crisis detection (Phillips and Shi, 2017; Phillips et al., 2015a,b) and is now widely used by academic researchers, central bank economists, and fiscal regulators. We illustrate the effectiveness of this procedure with applications to the S&P financial market and the European sovereign debt sector. These applications are implemented using the *psymonitor* R package (Phillips et al., 2018) developed in conjunction with this chapter.

Keywords: Bubbles, Crises, Real-time detection, Recursive evolving test

^{*}This chapter draws on several of our earlier works on bubble detection and real time monitoring of financial markets. Phillips acknowledges research support from the Kelly Fund, University of Auckland. Shi acknowledges funding support from the Australian Research Council (DP150101716). We thank Itamar Caspi, Yang Hu, and Les Oxley for their helpful comments and Lihong Chen for valuable research assistance.

1 Introduction

Speculative behavior and crises in the financial system can inflict serious harm on the real economy. Central banks, regulators, and policy makers therefore seek effective early warning devices of such episodes to assist in maintaining economic and financial stability. To meet the need for ongoing market surveillance, the recent literature on bubble detection has focused on real-time monitoring techniques rather than ex post identification strategies which were emphasized in earlier research (see [Gürkaynak, 2008](#) for a review).

A practical real-time bubble detection method was proposed by [Phillips et al. \(2015a,b\)](#), PSY hereafter and has now been successfully employed as an early warning alert system for exuberance in a wide variety of financial, commodity, and real estate markets. For many of these diverse applications readers may usefully refer to the following papers: [Bohl \(2003\)](#); [Etienne et al. \(2014a,b\)](#); [Gutierrez \(2012\)](#); [Pavlidis et al. \(2016\)](#); [Adämmer and Bohl \(2015\)](#); [Figuerola-Ferretti et al. \(2015\)](#); [Caspi et al. \(2015\)](#); [Caspi \(2016\)](#); [Shi et al. \(2016\)](#); [Phillips and Yu \(2011, 2013\)](#); [Greenaway-McGrey and Phillips \(2016\)](#); [Hu and Oxley \(2017a,b,c, 2018a,b\)](#). The potential of the PSY method has been recognized by central bank economists and fiscal regulators, as well as more widely in the financial industry and financial press. It is now employed by the Federal Reserve Bank of Dallas, providing an exuberance indicator for 23 international housing markets.^a Researchers from many central banks, including the Hong Kong Monetary Authority ([Yiu and Jin, 2013](#)), the Central Bank of Colombia ([Amador-Torres et al., 2018](#); [Gomez-Gonzalez et al., 2018](#)), and Bank of Israel ([Caspi, 2016](#)), have applied the PSY test to study real estate bubbles in their respective economies.

The PSY procedure serves as an early warning device for crises, as indicated in [Phillips and Shi \(2017\)](#). This capability has been noted in the many recent studies considering stock prices and exchange rates and other financial time series. See, [Phillips et al. \(2015a\)](#); [Phillips and Shi \(2017, 2018\)](#); [Shi \(2017\)](#); [Deng et al. \(2017\)](#); [Yiu and Jin \(2013\)](#); [Fantazzini \(2016\)](#); [Hu and Oxley \(2017b\)](#), among others.

The PSY procedure employs the augmented Dickey–Fuller (ADF) model specification and a recursive evolving algorithm. The recursive evolving algorithm relies only on historical information and permits a time-varying model structure. The method has general applications in regression. When applied to ADF regressions, the recursive evolving algorithm fixes the end point on the observation of interest and searches for the optimal starting point. As such, it minimizes the impact of previous episodes on the current identification and is less sensitive to the random choice of sample starting point. In effect, the method selects the most appropriate initialization for conducting a regression fit with given data, as considered in early work on econometric model determination ([Phillips, 1996](#)).

^aSee <https://www.dallasfed.org/institute/houseprice>.

In detecting change and for bubble identification the recursive evolving algorithm has been shown to outperform the forward recursive algorithm (Phillips et al., 2011), the rolling window approach (Chong and Hurn, 2017; Shi, 2007), and the cusum monitoring strategy (Homm and Breitung, 2012). Unlike regime switching methods (Hall et al., 1999; Shi and Song, 2016), it is a real-time procedure and easy to implement in practical work.

The identification of bubbles is based on their defining time series characteristics. During the expansionary phase of a bubble, asset prices follow a mildly explosive process as opposed to the martingale behavior that is typical during normal market conditions. In the event of a crisis or rapidly escalating credit risk, asset price (and hence bond yield) dynamics typically switch to a random drift martingale often accompanied by a large negative shock or a sequence of negative shocks. The PSY procedure which provides a joint test for the drift and the autoregressive coefficients of the ADF model is capable of detecting both bubbles and crises. The approach also delivers a mechanism for date stamping the origination and termination of bubbles. Consistency of the estimated bubble origination and termination dates was established in Phillips et al. (2015b) and Phillips and Shi (2018) under various data generating processes; and Phillips and Shi (2017) proved consistency of the estimated switch date for crises.

Harvey et al. (2016) showed that the presence of heteroskedasticity can affect the performance of the forward recursive method of Phillips et al. (2011) and can lead to severe size distortions in testing. The same fragility to heteroskedasticity is expected for the PSY procedure. Several methods have been proposed to overcome this problem (Harvey et al., 2016, 2018b). The wild bootstrap approach proposed by Harvey et al. (2016) has been found to have satisfactory asymptotic and finite sample performance. But an additional issue arises from the sequential nature of recursive hypothesis testing. It is well known that the probability of making false positive conclusions rises with the number of hypotheses tested, a phenomenon that is sometimes referred to as the *multiplicity* or *family-wise size control* issue in testing. This problem is common to all recursive testing procedures. In this chapter we propose a new bootstrap procedure that simultaneously addresses both heteroskedasticity and multiplicity issues in testing.

The PSY procedure is now a standard item in the econometric toolkit. Matlab, Eviews, and R software programs are available for practical implementation.^b This chapter illustrates implementation of the methodology with a new R package that incorporates the bootstrap procedure for dealing with heteroskedasticity and multiplicity in recursive testing. With this software we apply the procedure to S&P 500 stock market data to detect both bubble and crisis episodes in the stock market and to the European sovereign debt

^bSee the website <https://sites.google.com/site/shupingshi/home/codes> for the Matlab codes, the *Rtadf* Eviews Addin (Caspi, 2017), and the *MultipleBubbles* (Araujo et al., 2018) and *exuber* (Vasilopoulos et al., 2018) packages in R.

market to detect episodes of escalating credit risk. The new R package ([Phillips et al., 2018](#)) is named *psymonitor* and can be installed with the following command sequence:

```
install.packages("psymonitor")
library(psymonitor)
```

The rest of the chapter is organized as follows: [Section 2](#) introduces the PSY procedure. The rationale and limiting properties of the PSY procedure for bubble identification (crisis detection) are described and illustrated in [Section 3](#) ([Section 4](#)). [Section 5](#) introduces the new bootstrap procedure for accommodating heteroskedasticity and addressing multiplicity issues. Empirical applications to the S&P 500 market and the European sovereign market are given in [Section 6](#). [Section 7](#) concludes.

2 The PSY Procedure

The PSY procedure was originally designed to identify and date stamp explosive periods in asset prices. Subsequent research ([Phillips and Shi, 2017](#)) has shown that the method has detective power against both speculative bubbles and market collapses, including flash crashes. The method is based on an ADF model specification for the fitted regression equation but uses flexible window widths in its implementation to take time-varying dynamics and structural breaks into consideration.

2.1 The Augmented Dickey–Fuller test

It is well known in the unit root literature that the limit distribution of the ADF statistic depends on both the null hypothesis and the precise regression model specification.^c Appropriate choices of both therefore have a material impact in practical implementation.

The null hypothesis (H_0) of the PSY test captures normal market behaviors and states that asset prices follow a martingale process with a mild drift function such that ([Phillips et al., 2014](#))

$$y_t = g_T + y_{t-1} + u_t, \quad (1)$$

where $g_T = kT^{-\gamma}$ (with constant k , $\gamma > 1/2$, and sample size T) captures any mild drift that may be present in prices but which is of smaller order than the martingale component and is therefore asymptotically negligible.

The regression model chosen for the PSY procedure is

$$\Delta y_t = \mu + \rho y_{t-1} + \sum_{j=1}^p \phi_j \Delta y_{t-j} + v_t, \quad (2)$$

^cSee [Hamilton \(1994\)](#) for a textbook discussion and [Phillips et al. \(2014\)](#) for details in the context of bubble testing with localized drift specifications.

where for implementation purposes the regression error v_t is assumed to satisfy $v_t \stackrel{i.i.d.}{\sim} (0, \sigma^2)$. The p lag terms of Δy_t are included to take care of potential serial correlation. The lag order p is often selected by information criteria. The regression model includes an intercept but no time trend and nests the null hypothesis as a special case with $\mu = g_T$ and $\rho = 0$. The ADF statistic is simply the t -ratio of the least squares estimate of the coefficient ρ .

The *i.i.d* error condition may be replaced with a martingale difference sequence (mds) condition in (2). More general specifications on the error u_t in the generating mechanism (1), such as those in [Assumption 1](#) below, may be employed and are accommodated by allowing the regression lag order $p \rightarrow \infty$ as $T \rightarrow \infty$ in (2). Nonparametric adjustments for serial correlation may also be used, such as those developed in [Phillips \(1987\)](#) and [Phillips and Perron \(1988\)](#).

Assumption 1. *The error term u_t is allowed to be serial correlated such that*

$$u_t = \psi(L)\varepsilon_t = \sum_{j=0}^{\infty} \varphi_j \varepsilon_{t-j},$$

where $\sum_{j=0}^{\infty} |\varphi_j| < \infty$ and ε_t is an mds satisfying:

- (i) ε_t is strongly uniformly integrable with a dominating random variable η that satisfies $\mathbb{E}(\eta^2 \ln^+ |\eta|) < \infty$;
- (ii) $T^{-1} \sum_{t=1}^T \mathbb{E}(\varepsilon_t^2 | \mathcal{F}_{t-1}) \rightarrow a.s. \sigma^2$, where $\mathcal{F}_t = \sigma\{\varepsilon_t, \varepsilon_{t-1}, \dots\}$ is the natural filtration.

Under [Assumption 1](#) ε_t is potentially conditionally heteroskedastic, as for instance under stable ARCH or GARCH errors. The partial sums of ε_t satisfy the functional law ([Phillips and Solo, 1992](#))

$$T^{-1/2} \sum_{t=1}^{\lfloor Tr \rfloor} \varepsilon_t \Rightarrow \sigma W(r), \quad (3)$$

where W is standard Brownian motion, \Rightarrow signifies weak convergence on the Skorohod space $D[0, 1]$, and $\lfloor \cdot \rfloor$ signifies the integer part of the argument.

Under the null hypothesis (1), [Assumption 1](#), and regression (2) with side conditions that ensure $p \rightarrow \infty$, the ADF statistic has a limit distribution given by [Phillips et al. \(2014\)](#)

$$ADF \Rightarrow \frac{\frac{1}{2} [W(1)^2 - 1] - W(1) \int_0^1 W(s) ds}{\left[\int_0^1 W(s)^2 ds - \left(\int_0^1 W(s) ds \right)^2 \right]^{1/2}},$$

where \Rightarrow denotes convergence in distribution on \mathbb{R} .

2.2 The Recursive Evolving Algorithm

The recursive evolving algorithm of PSY enables real-time identification of bubbles and crises while allowing for the presence of multiple structural breaks within the sample period. Phillips et al. (2015a,b) show that this algorithm is superior to the forward expanding and rolling window algorithms in bubble identification, especially when the sample period contains multiple bubbles.

For the convenience of exposition, we use the standard “fraction of the total sample” notation for observations. Thus if $t = \lfloor Tr \rfloor$ is the integer part of $T r$, then observation t is represented fractionally as observation r and then the total sample runs over values of r from 0 to 1. Suppose the observation of interest is r^\dagger . The PSY procedure calculates the ADF statistic recursively from a backward expanding sample sequence. Let r_1 and r_2 be the start and end points of the regression sample. The ADF statistic calculated from this sample is denoted by $ADF_{r_1}^{r_2}$. We fix the end point of all samples on the observation of interest such that $r_2 = r^\dagger$ and allow the start point r_1 to vary within its feasible range, i.e., $[0, r^\dagger - r_0]$, where r_0 is the minimum window required to initiate the regression. The recommended setting of r_0 for practical implementation is $r_0 = 0.01 + 1.8/\sqrt{T}$. The PSY statistic is the supremum taken over the values of all the ADF statistics in the entire recursion, which is represented mathematically as

$$PSY_{r^\dagger}(r_0) = \sup_{r_1 \in [0, r^\dagger - r_0], r_2 = r^\dagger} \{ ADF_{r_1}^{r_2} \}.$$

The supremum enables the selection of the “optimal” starting point of the regression in the sense of providing the largest ADF statistic.

The PSY test can be conducted for each individual observation of interest ranging from r_0 to 1, i.e., for $r^\dagger \in [r_0, 1]$. The recursive calculation evolves as the observation of interest moves forward and therefore the procedure is called a recursive evolving algorithm. See Fig. 1 for a graphical illustration of the algorithm. The corresponding PSY statistic sequence is $\{PSY_{r^\dagger}(r_0)\}_{r^\dagger \in [r_0, 1]}$.

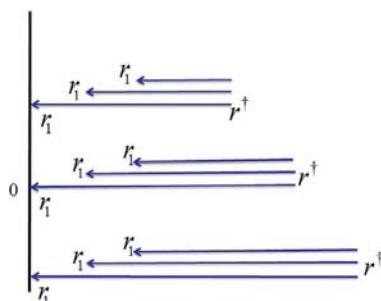


FIG. 1 The recursive evolving algorithm with $r_1 \in [0, r^\dagger - r_0]$ and $r_2 = r^\dagger$.

Calculation of the PSY statistic sequence can be achieved with the command *PSY* contained in the *psymonitor* R package. This routine requires the input of data (*y*), a minimum window size (*swindow0*), and a choice of information criterion for the lag order selection (*IC* and *adf lag*). The syntax of the call is.

PSY(y, swindow0, IC, adflag).

IC has value 0 when a fixed lag order of *adflag* is used, 1 for use of an AIC lag order selector, and 2 for a BIC order selector. In the latter two cases, a maximum lag order *adflag* is employed in the information criteria.

Under the null hypothesis of normal market conditions and the conditions described earlier, the PSY statistic has the following limit distribution (Phillips et al., 2015a)

$$\sup_{r_1 \in [0, r^* - r_0], r_2 = r^*} \left\{ \frac{\frac{1}{2} r_w \left[W(r_2)^2 - W(r_1)^2 - r_w \right] - \int_{r_1}^{r_2} W(s) ds [W(r_2) - W(r_1)]}{r_w^{1/2} \left[r_w \int_{r_1}^{r_2} W(s)^2 ds - \left(\int_{r_1}^{r_2} W(s) ds \right)^2 \right]^{1/2}} \right\}, \quad (4)$$

where $r_w = r_2 - r_1$.

The origination of a bubble or crisis episode is taken to be where the PSY test statistic first exceeds its critical value—a first stopping time for this episode. Likewise, the termination date is taken to be where the supremum test statistic subsequently falls below its critical value—a second stopping time for this episode. Suppose the sample contains only one episode originating at r_e and finishing at r_f . The estimated origination and termination dates (denoted by \hat{r}_e and \hat{r}_f) are then given by the stopping times

$$\hat{r}_e = \inf_{r^* \in [r_0, 1]} \{r^* : PSY_{r^*}(r_0) > cv_{r^*}(\beta_T)\}, \quad (5)$$

$$\hat{r}_f = \inf_{r^* \in [\hat{r}_e, 1]} \{r^* : PSY_{r^*}(r_0) < cv_{r^*}(\beta_T)\}, \quad (6)$$

where $cv_{r^*}(\beta_T)$ is the 100 ($1 - \beta_T$) critical value (quantile of the distribution) of the $PSY_{r^*}(r_0)$ statistic. The notation for test size β_T being sample size dependent allows for the property that $\beta_T \rightarrow 0$ as $T \rightarrow \infty$. This property in turn leads to $cv_{r^*}(\beta_T) \rightarrow \infty$ under the null hypothesis, thereby ensuring that the probability of falsely detecting the presence of a bubble under the null passes to zero in large samples.

Estimation of the origination and termination dates are achieved by the *locate* function in the R package.

locate(ind, date),

where ind is the vector of PSY indicators taking value one when the test statistic is above the critical value and zero otherwise and $date$ is the vector of calendar dates associated with the observation.

3 The PSY Test for Bubble Identification

3.1 The Rationale

To illustrate the idea of bubble identification, consider the present value asset price formula

$$P_t = \sum_{i=0}^{\infty} \left(\frac{1}{1+r_f} \right)^i \mathbb{E}_t(D_{t+i}) + B_t, \quad (7)$$

where P_t is the price of the asset, D_t is the payoff received from the asset, r_f is the risk-free interest rate, $\mathbb{E}_t(\cdot)$ is the conditional expectation operator given information to time t , and B_t is the bubble component. The bubble component satisfies the submartingale property (Diba and Grossman, 1988)

$$\mathbb{E}_t(B_{t+1}) = (1+r_f)B_t. \quad (8)$$

In the absence of a bubble, the degree of nonstationarity of the asset price is controlled entirely by the dividend series and hence is believed from empirical evidence to be at most $I(1)$. On the other hand, asset prices will be explosive in the presence of a bubble component in formula (7) whenever the initialization $B_0 > 0$ in (8).

Asset price dynamics over the expansionary phase of a bubble period may be modeled in terms of a mildly explosive process (Phillips et al., 2011; Phillips and Magdalinos, 2007; Phillips and Yu, 2009) of the form

$$\log P_t = \delta_T \log P_{t-1} + u_t, \quad (9)$$

where the autoregressive coefficient $\delta_T = 1 + cT^{-\eta}$ mildly exceeds unity (with $c > 0$ and $\eta \in (0, 1)$) and yet still lies in its general vicinity. Detection of a bubble process in the data is therefore equivalent to distinguishing a martingale process of asset prices from a mildly explosive process. This can be achieved by the PSY procedure with null and alternative hypotheses specified as

$$\begin{aligned} H_0 : \mu &= g_T \text{ and } \rho = 0, \\ H_A : \mu &= 0 \text{ and } \rho > 0. \end{aligned}$$

3.2 Consistency

The data generating process (9) assumes the presence of an expansionary bubble over the entire sample period. In practice, bubbles exist only over subperiods and involve periods of collapse or contraction as well as expansion, thereby justifying the terminology. A given sample of data may include only

martingale behavior, a single bubble episode set amidst martingale behavior on either side, or a sequence multiple bubble episodes interspersed amidst normal martingale behavior. An important task in the real-time dating literature is to demonstrate consistency of the estimated origination and termination dates of such bubble episodes.

The simplest example is a sample which contains a single bubble expansionary episode which does not terminate or collapse within the sample period. Specifically, asset prices follow a small drift martingale as in (1) before period $\tau_e = \lfloor r_e T \rfloor$ and then switch to a mildly explosive process as in (9), viz.,

$$\log P_t = \begin{cases} g_T + \log P_{t-1} + u_t & \text{if } t < \tau_e \\ \delta_T \log P_{t-1} + u_t & \text{if } t \geq \tau_e \end{cases} \quad (10)$$

This DGP can be extended to include the bubble collapse dynamics. Various patterns of collapse have been considered in the literature. Suppose the end date of the bubble episode is $\tau_f = \lfloor r_f T \rfloor$. [Phillips et al. \(2011\)](#) proposed an abrupt bubble collapse pattern where asset prices return immediately to the level before the bubble origination allowing for a stationary perturbation, so that

$$\log P_{\tau_f} = \log P_{\tau_e-1} + O_p(1). \quad (11)$$

[Phillips and Shi \(2018\)](#) recommended a mildly integrated reversion pattern for observations in the collapsing regime in which prices follow the mechanism

$$\log P_t = \gamma_T \log P_{t-1} + u_t, \quad (12)$$

where the autoregressive coefficient $\gamma_T = 1 - c_1 T^{-\beta}$ is smaller than unity ($c_1 < 0$ and $\beta \in (0, 1)$). By varying the value of β , this process can generate abrupt, randomly disturbed, or smooth patterns of collapse behavior.

Under the data generating process (10), [Phillips et al. \(2015b\)](#) show that the PSY test statistic has order of magnitude $O_p(1)$ if the observation of interest r^* falls in the normal regime and diverges to positive infinity at the rate $O_p(T^{1/2} \delta_T^{\tau^* - \tau_e})$ with $\tau^* = \lfloor r^* T \rfloor$ if the observation lies in the bubble regime. For observations in the collapse regime, the PSY statistic diverges to negative infinity at the rate $O_p(T^{(1-\eta)/2})$ when the bubble collapses in the fashion of (11) or $O_p(T^{\omega(\eta, \beta)})$ ^d when the collapse process is (12). It transpires that under the condition that test size $\beta_T \rightarrow 0$ as $T \rightarrow \infty$ and

$$\frac{1}{cv_{r^*}(\beta_T)} + \frac{cv_{r^*}(\beta_T)}{T^{1/2} \delta_T^{\tau^* - \tau_e}} \rightarrow 0,$$

we have the consistency of the estimated bubble origination and termination dates, i.e., $\hat{r}_e \rightarrow r_e$ and $\hat{r}_f \rightarrow r_f$.

^d $\omega(\cdot)$ is a linear function of the arguments.

The process can be generalized to allow for the presence of multiple bubbles. Consistency of the estimated bubble origination and termination dates in the presence of multiple bubbles was shown by Phillips et al. (2015b) for the DGP with the abrupt collapsing pattern (11) and by Phillips and Shi (2018) with the mildly integrated reverting pattern (12).

4 The PSY Test for Crisis Identification

4.1 The Rationale

Market crashes are defined as a discontinuity in asset prices that is characterized by large downward movements (Barlevy and Veronesi, 2003; Gennotte and Leland, 1990). The dynamics of asset prices during crisis periods may be modeled as a random drift martingale process (Phillips and Shi, 2018)

$$\log P_t = -L_t + \log P_{t-1} + u_t, \quad (13)$$

in which L_t is a random sequence independent of u_t . The sequence L_t produces a random drift in the observed price process and L_t may take various forms, which lead to a corresponding variety of collapse mechanism. The simple process used in Phillips and Shi (2018) follows an asymmetric scaled uniform distribution such that

$$L_t = L b_t, \quad b_t \stackrel{i.i.d.}{\sim} U[-\epsilon, 1], \quad 0 < \epsilon < 1.$$

where L is a positive scale quantity measuring shock intensity and b_t is uniform on an interval ranging from a (usually small) negative value $-\epsilon$ to unity. The mean of the drift term takes a negative value (i.e., $-(1 - \epsilon)L/2$) and hence the process exhibits an overall downward trend. The magnitude of this downward trend depends on the values of the scalar parameters L and ϵ .

Suppose P_t is the price of a stock and the logarithmic dividend is a martingale with drift generated as

$$\log D_t = \alpha + \log D_{t-1} + v_t, \quad (14)$$

where α is a constant and the v_t are mds innovations. Under the price process (13), the logarithmic price-dividend ratio also follows a random drift martingale process of the form

$$\log P_t/D_t = -L_t^* + \log P_{t-1}/D_{t-1} + \varepsilon_t^*, \quad (15)$$

where $L_t^* = L b_t^*$ with $b_t^* \stackrel{i.i.d.}{\sim} U[-\epsilon + \alpha/L, 1 + \alpha/L]$ and $\varepsilon_t^* = \varepsilon_t - v_t$.

Suppose P_t is the price of a τ -period discount bond. The relationship between the continuously compounded zero-coupon nominal yield to maturity (z_t) and the bond price is

$$z_t = -\frac{\log P_t}{\tau}.$$

Bond yields often serve as a proxy for credit risk. A loan default or other credit events may trigger a sharp decline in bond prices and hence a fast expansion in bond yields. Under the assumption of a bond price crash (13), bond yields follow

$$z_t = \frac{1}{\tau} L_t + z_{t-1} - \frac{u_t}{\tau}. \quad (16)$$

The drift term has a positive mean of $(1 - \epsilon) L/(2\tau)$, implying an overall upward trend in the dynamics.

In the setting of this model detecting financial crises or ballooning credit risk is equivalent to distinguishing a martingale process with a small drift (null) from a random-drift martingale process (alternative). The null and alternative hypotheses of the PSY test for crises may now be formulated in terms of the fitted ADF regression Eq. (2) as follows:

$$\begin{aligned} H_0 &: \mu = g_T \text{ and } \rho = 0 \\ H_{1,\text{crash}} &: \mu = K \text{ and } \rho = 0. \end{aligned}$$

where K is the expected value of the random drift process L_t and g_T is an asymptotically negligible deterministic drift as in (1).

4.2 Consistency

The specification (13) can be modified to switch on or off depending on the financial environment. The data generating process for asset prices considered in Phillips and Shi (2017) is

$$\log P_t = \begin{cases} g_T + \log P_{t-1} + u_t & \text{if } t < \tau_e \\ -L_t + \log P_{t-1} + u_t & \text{if } t \geq \tau_e \end{cases}. \quad (17)$$

The origination of the event is denoted again by τ_e . Suppose that P_t is the price of a discount bond. It is straightforward to show that the bond yield z_t follows a stochastic process that switches between a martingale with a small deterministic drift and a martingale with a positively scaled random drift

$$z_t = \begin{cases} -\frac{1}{\tau} g_T + z_{t-1} - \frac{u_t}{\tau} & \text{if } t < \tau_e \\ \frac{1}{\tau} L_t + z_{t-1} - \frac{u_t}{\tau} & \text{if } t \geq \tau_e \end{cases}. \quad (18)$$

Phillips and Shi (2017) show that under the DGP (17), the PSY test statistic diverges to positive infinity at the rate $O_p(T^{1/2})$ as the test proceeds from the normal regime to the crash regime. It follows that the PSY procedure can consistently estimate the break date τ_e when

$$\frac{1}{cv_{r^\dagger}(\beta_T)} + \frac{cv_{r^\dagger}(\beta_T)}{T^{1/2}} \rightarrow 0.$$

5 A New Composite Bootstrap

The bootstrap procedure described here combines the two procedures of [Harvey et al. \(2016\)](#) and [Shi et al. \(2018\)](#). It is designed to mitigate the potential influence of unconditional heteroskedasticity and to address the multiplicity issue in recursive testing. Let $\tau_0 = \lfloor Tr_0 \rfloor$ and τ_b be the number of observations in the window over which size is to be controlled.

Step 1: Using the full sample period, estimate the regression model (2) under the imposition of the null hypothesis of $\rho=0$ and obtain the estimated residual e_r .

Step 2: For a sample size $\tau_0 + \tau_b - 1$, generate a bootstrap sample given by

$$\Delta y_t^b = \sum_{j=1}^p \hat{\phi}_j \Delta y_{t-j}^b + e_t^b \quad (19)$$

with initial values $y_i^b = y_i$ with $i = 1, \dots, j+1$, and where the $\hat{\phi}_j$ are the OLS estimates obtained in the fitted regression from Step 1. The residuals $e_t^b = w_t e_t$ where w_t is randomly drawn from the standard normal distribution and e_t is randomly drawn with replacement from the estimated residuals e_r .

Step 3: Using the bootstrapped series, compute the PSY test statistic sequence $\{PSY_t^b\}_{t=\tau_0}^{\tau_0+\tau_b-1}$ and the maximum value of this test statistic sequence, giving

$$\mathcal{M}_t^b = \max_{t \in [\tau_0, \tau_0 + \tau_b - 1]} (PSY_t^b).$$

Step 4: Repeat Steps 2–3 for $B = 999$ times.

Step 5: The critical value of the PSY procedure is now given by the 95% percentile of the $\{\mathcal{M}_t^b\}_{b=1}^B$ sequence.

Step 2 of this iteration implements a wild bootstrap to address heteroskedasticity; and Steps 3–5 of the iteration replicate the PSY recursive test sequence and create critical values that account for multiplicity in the test sequence recursion.

The bootstrap procedure can be implemented with the following call syntax in R with the package.

```
cvPSYwmbot(y, swindow0, IC, adflag, Tb, nboot, nCores),
```

where the argument Tb corresponds to τ_b , $nboot$ is the number of bootstrap repetitions, and $nCores$ is the number of cores used for the calculation. The other arguments are the same as those described earlier.

6 Empirical Applications with R

6.1 Example 1: The S&P 500 Market

The S&P 500 stock market has been a central focus of attention in global financial markets due to the size of this market and its impact on other

financial markets. As an illustration of the methods discussed in this chapter, we conduct a pseudo real-time monitoring exercise for bubbles and crises in this market with the PSY strategy. The sample period runs from January 1973 to July 2018, downloaded monthly from *Datstream International*. The price-dividend ratios are computed as the inverse of dividend yields. The first step is to import the data to R, using the following code:

```
sp500 <- read.csv("sp500.csv")
date <- as.Date(sp500[,1], "%d/%m/%Y")
dy <- sp500[,2]
pd <- 1/dy
```

In the presence of a speculative bubble, asset prices characteristically deviate in an explosive way from fundamentals, representing exuberance in the speculative behavior driving the market. In the present case, this deviation implies that the log price-dividend ratio is expected to follow an explosive process over the expansive phase of the bubble. But during crisis periods, the price-dividend ratio is expected to follow a random (downward) drift martingale process, in contrast to a small (local to zero) constant drift martingale process that typically applies under normal market conditions. According to the theory detailed in Sections 3 and 4, we expect to witness rejection of the null hypothesis in the PSY test empirical outcomes during both bubble and crisis periods.

Fig. 2 plots the price-to-dividend ratio of the S&P 500 index. We observe a dramatic increase in the data series in the late 1990s, followed by a rapid fall in the early 2000s. The market experienced another episode of slump in late 2008. With a training period of 47 observations, we start the pseudo real-time monitoring exercise from November 1976 onwards. The PSY test statistics are compared

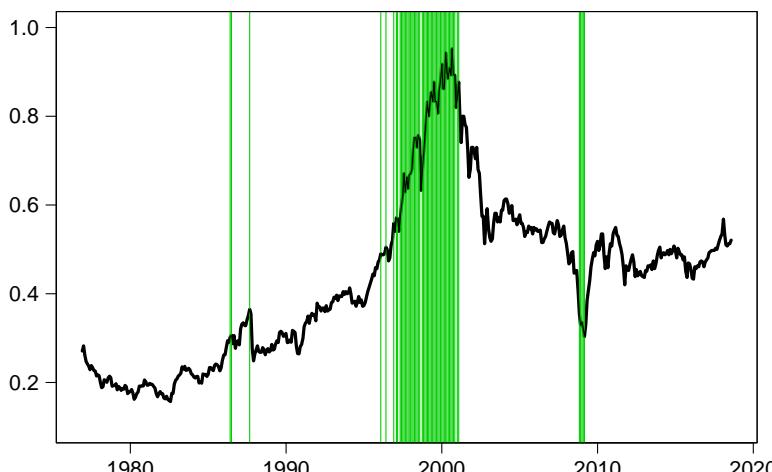


FIG. 2 Bubble and crisis periods in the S&P 500 stock market. The *solid line* is the price-to-dividend ratio and the *shaded areas* are the periods where the PSY statistic exceeds its 95% bootstrapped critical value.

with the 95% bootstrapped critical value. The empirical size is controlled over a 2-year period, i.e., by taking $\tau_b=24$. The lag order is selected by BIC with a maximum lag order of 6, applied to each subsample. The PSY statistic sequence and the corresponding bootstrap critical values can be calculated as follows in R.

```

y<-pd
obs<-length(y)
r0<-0.01+1.8/sqrt(obs)
swindow0<-floor(r0*obs)
dim<-obs-swindow0+1

IC<-2
adflag<-6 yr<-2
Tb<-12*yr+swindow0-1
nboot<-999
nCore<-2

bsadf<-PSY(y,swindow0,IC,adflag)
quantilesBsadf<-cvPSYwmboot(y,swindow0,IC,adflag,Tb,nboot,nCore)

```

The identified origination and termination dates can be calculated and viewed with the following commands.

```

date<-date[swindow0:obs]
quantile95<-quantilesBsadf%*%matrix(1,nrow=1,ncol=dim)
ind95<-(bsadf>t(quantile95[2,]))*1

OT <-locate(ind95,date)
BCdates<-disp(OT,obs)
print(BCdates)

```

where the last two command syntax print the dates on the screen with the first (second) column being the origination (termination) date. The outputs are.

| | Start | End |
|---|------------|------------|
| 1 | 1986-05-30 | 1986-06-30 |
| 2 | 1987-07-31 | 1987-08-31 |
| 3 | 1996-01-31 | 1996-01-31 |
| 4 | 1996-05-31 | 1996-05-31 |
| 5 | 1996-11-29 | 1997-02-28 |
| 6 | 1997-04-30 | 1998-07-31 |
| 7 | 1998-09-30 | 2000-10-31 |
| 8 | 2000-12-29 | 2001-01-31 |
| 9 | 2008-10-31 | 2009-02-27 |

The identified periods are shaded in green in Fig. 2. As is evident in the figure, the procedure detects two bubble episode and one crisis episode. The first bubble episode only lasts for 3 months (1986M05–M06 and 1987M08) and occurred before the Black Monday crash on October 1987. The second bubble episode is the well-known dot-com bubble, starting from January 1996 and terminating in October 2000 (with several breaks in between). For the dot-com bubble episode the identified starting date for market exuberance occurs well before the speech of the former chairman of the Federal Reserve Bank Alan Greenspan in December 1996 where the now famous question “how do we know when irrational exuberance has unduly escalated asset values” was posed to the audience and financial world. The identified subprime mortgage crisis starts in October 2008, which is 1 month after the collapse of Lehman Brothers, and terminates in February 2009.

The codes for generating the plot and shaded overlays in the figure are as follows:

```
plot(date,y[swindow0:obs],xlim=c(min(date),max(date)),
      ylim=c(0.1,1),
      xlab=",ylab=",type='l',lwd=3)
for(i in 1:length(date)){
  if (ind95[i]==1){abline(v=date[i],col=3)}
points(date,y[swindow0:obs],type='l')
box(lty=1)
dev.off()
```

6.2 Example 2: Credit Risk in the European Sovereign Sector

The European sovereign debt sector experienced an extremely turbulent period over the last decade, which caused significant harm to the real economy (Acharya et al., 2018) and led to an unprecedented level of unemployment (Karafolas and Alexandrakis, 2015). The PSY detection algorithm can serve as a useful early warning mechanism for escalating credit risk, which is acknowledged as a leading indicator of financial and economic crises, and thereby enable timely policy action and effective risk management to avert more serious economic damage. To show the potential efficacy of this early warning system, we conduct a pseudo monitoring exercise of credit risk in the European sovereign sector.

Credit risk in the European sovereign sector is proxied by an index constructed as a GDP weighted 10-year government bond yield of the GIIPS (Greece, Ireland, Italy, Portugal, and Spain) countries.^e The PSY strategy is applied to the spread between the GIIPS bond yield index and the 10-year government bond yield of Germany (used as a proxy for a prevailing

^eThese are the five EU countries that were unable to refinance their government debt or to bail out banks on their own during the debt crisis.

benchmark of economic fundamentals). The sample data runs from June 1997 to June 2016 and was downloaded from *Datstream International*. The GDP data are downloaded quarterly and converted to a monthly frequency by assuming a constant value within each quarter.

```
data <- read.csv("spread.csv")
date <- as.Date(data[,1], "%d/%m/%Y")
spread <- data[,2]
y<-spread
```

Fig. 3 plots the bond yield spread over the sample period. The bond yield index experienced a rapid and substantial rise between 2008 and 2009. It continued to mount to historical highs from 2010 onwards and peaked in June 2012. The bond yield index has dropped since then and becomes relatively stable over the last 2 years. The codes for implementing the PSY procedure are identical to those for Example 1. The estimated start and end dates of the crisis episodes are displayed below.

| | Start | End |
|---|------------|------------|
| 1 | 2008-03-23 | 2008-03-23 |
| 2 | 2008-10-23 | 2009-03-23 |
| 3 | 2010-05-23 | 2012-08-23 |

The shaded areas in **Fig. 3** are the identified periods of crisis obtained using the 95% bootstrap critical values. The first alarm signal of risk appeared

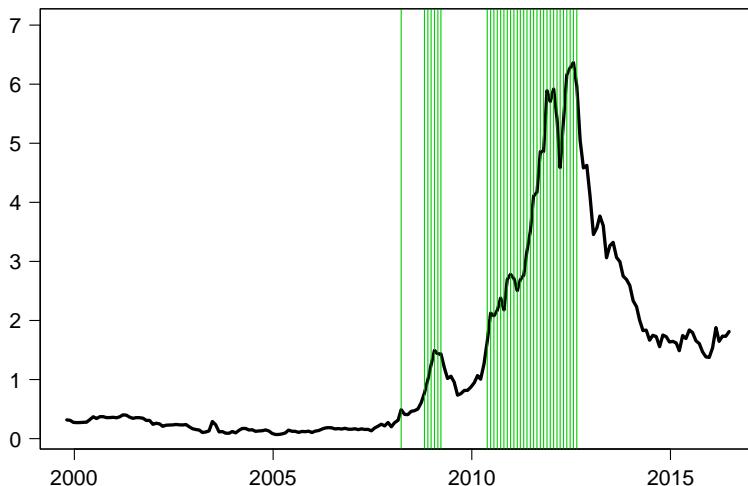


FIG. 3 Crisis episodes in the European sovereign sector. The *solid line* is the 10-year government bond yield spread between the GIIPS countries and Germany and the shaded areas are the periods where the PSY statistic exceeds its 95% bootstrap critical value.

in March 2008 and lasts for 1 month. The alarm was triggered again after the collapse of Lehman Brothers in October 2008 and turns off in March 2009. The stress indicator switched on again in May 2010 and lasted until August 2012.

7 Conclusion

The recursive evolving test algorithm proposed by Phillips et al. (2015a,b) provides a real-time empirical device for detecting speculative bubbles, crises, and ballooning credit risks that can foreshadow impending damage to the real economy. The multifunctionality and the real-time features of this algorithm assist policymakers in market surveillance and investors in risk management. The approach has enjoyed widespread use in academic circles and among central bank economists.

This chapter overviews the main features of the PSY approach and details a new combined bootstrap procedure for dealing with both heteroskedasticity and multiplicity issues in recursive inference methods. A new R package *psymonitor* (complete with the combined bootstrap procedure) is provided for convenient implementation of the methods. For empirical illustration of the use of the R codes, the procedures are applied to the S&P 500 stock market for the detection of bubbles and crises and to the European sovereign debt sector for detection of ballooning credit risks. We hope that the R package and these illustrations^f will assist in making these methods widely available to empirical researchers, industry economists, and policy makers.

References

- Acharya, V.V., Eisert, T., Eufinger, C., Hirsch, C., 2018. Real effects of the sovereign debt crisis in Europe: evidence from syndicated loans. *Rev. Finance Stud.* 31 (8), 2855–2896.
- Adämmer, P., Bohl, M.T., 2015. Speculative bubbles in agricultural prices. *Q. Rev. Econ. Finance* 55, 67–76.
- Amador-Torres, J.S., Gomez-Gonzalez, J.E., Sanin-Restrepo, S., 2018. Determinants of housing bubbles' duration in OECD countries. *Int. Finance*.
- Araujo, P., Lacerda, G., Phillips, P.C.B., Shi, S., 2018. Test and Detection of Explosive Behaviors for Time Series. R Foundation for Statistical Computing, Vienna, Austria. <https://cran.r-project.org/web/packages/MultipleBubbles/>.
- Barlevy, G., Veronesi, P., 2003. Rational panics and stock market crashes. *J. Econ. Theory* 110 (2), 234–263.
- Bohl, M.T., 2003. Periodically collapsing bubbles in the US stock market? *Int. Rev. Econ. Finance* 12 (3), 385–397.

^fFor more illustrations, see <https://itamarcaspi.github.io/psymonitor/>.

- Caspi, I., 2016. Testing for a housing bubble at the national and regional level: the case of Israel. *Empir. Econ.* 51 (2), 483–516.
- Caspi, I., 2017. Rtadf: testing for bubbles with EViews. *J. Stat. Softw.* 81 (1), 1–16.
- Caspi, I., Katzke, N., Gupta, R., 2015. Date stamping historical periods of oil price explosivity. *Energy Econ.* 70, 1876–2014.
- Chong, J., Hurn, S., 2017. Testing for Speculative Bubbles: Revisiting the Rolling Window. Queensland University of Technology. Working paper.
- Deng, Y., Girardin, E., Joyeux, R., Shi, S., 2017. Did bubbles migrate from the stock to the housing market in China between 2005 and 2010? *Pac. Econ. Rev.* 22 (3), 276–292.
- Diba, B.T., Grossman, H.I., 1988. Explosive rational bubbles in stock prices? *Am. Econ. Rev.* 78 (3), 520–530.
- Etienne, X.L., Irwin, S.H., Garcia, P., 2014a. Bubbles in food commodity markets: four decades of evidence. *J. Int. Money Finance* 42, 129–155.
- Etienne, X.L., Irwin, S.H., Garcia, P., 2014b. Price explosiveness, speculation, and grain futures prices. *Am. J. Agric. Econ.* 97 (1), 65–87.
- Fantazzini, D., 2016. The oil price crash in 2014/15: was there a (negative) financial bubble? *Energy Policy* 96, 383–396.
- Figuerola-Ferretti, I., Gilbert, C.L., McCrorie, J.R., 2015. Testing for mild explosivity and bubbles in LME non-ferrous metals prices. *J. Time Ser. Anal.* 36 (5), 763–782.
- Gennotte, G., Leland, H., 1990. Market liquidity, hedging, and crashes. *Am. Econ. Rev.* 80 (5), 999–1021.
- Gomez-Gonzalez, J.E., Gamboa-Arbeláez, J., Hirs-Garzón, J., Pinchao-Rosero, A., 2018. When bubble meets bubble: contagion in OECD countries. *J. Real Estate Finance Econ.* 56 (4), 546–566.
- Greenaway-McGrevy, R., Phillips, P.C.B., 2016. Hot property in New Zealand: empirical evidence of housing bubbles in the metropolitan centres. *N. Z. Econ. Pap.* 50 (1), 88–113.
- Gürkaynak, R.S., 2008. Econometric tests of asset price bubbles: taking stock. *J. Econ. Surv.* 22 (1), 166–186.
- Gutierrez, L., 2012. Speculative bubbles in agricultural commodity markets. *Eur. Rev. Agric. Econ.* 40 (2), 217–238.
- Hall, S.G., Psaradakis, Z., Sola, M., 1999. Detecting periodically collapsing bubbles: a Markov-switching unit root test. *J. Appl. Econom.* 14, 143–154.
- Hamilton, J.D., 1994. *Time Series Analysis*, first ed. Princeton University Press.
- Harvey, D.I., Leybourne, S.J., Sollis, R., Taylor, A.R., 2016. Tests for explosive financial bubbles in the presence of non-stationary volatility. *J. Empir. Finance* 38, 548–574.
- Harvey, D.I., Leybourne, S.J., Zu, Y., 2018b. Testing explosive bubbles with time-varying volatility. *Econ. Rev.* <https://doi.org/10.1080/07474938.2018.1536099>.
- Homm, U., Breitung, J., 2012. Testing for speculative bubbles in stock markets: a comparison of alternative methods. *J. Finance Econom.* 10 (1), 198–231.
- Hu, Y., Oxley, L., 2017a. Are there bubbles in exchange rates? some new evidence from G10 and emerging market economies. *Econ. Model.* 64, 419–442.
- Hu, Y., Oxley, L., 2017b. Bubble Contagion: Evidence From Japan's Asset Price Bubble of the 1980–90s. vol. 50, Technical report, 89–95.
- Hu, Y., Oxley, L., 2017c. Exuberance, Bubbles or Froth? Some Historical Results Using Long Run House Price Data for Amsterdam. University of Waikato, Norway and Paris. Technical Report, Working Paper.
- Hu, Y., Oxley, L., 2018a. Bubbles in US regional house prices: evidence from house price-income ratios at the state level. *Appl. Econ.* 50 (29), 3196–3229.

- Hu, Y., Oxley, L., 2018b. Do 18th century ‘bubbles’ survive the scrutiny of 21st century time series econometrics? *Econ. Lett.* 162, 131–134.
- Karafolas, S., Alexandrakis, A., 2015. Unemployment effects of the Greek crisis: a regional examination. *Proc. Econ. Finance* 19, 82–90.
- Pavlidis, E., Yusupova, A., Paya, I., Peel, D., Martínez-García, E., Mack, A., Grossman, V., 2016. Episodes of exuberance in housing markets: in search of the smoking gun. *J. Real Estate Finance Econ.* 53 (4), 419–449.
- Phillips, P.C.B., 1987. Time series regression with a unit root. *Econometrica* 55, 277–301.
- Phillips, P.C.B., 1996. Econometric model determination. *Econometrica* 64, 763–812.
- Phillips, P.C.B., Magdalinos, T., 2007. Limit theory for moderate deviations from a unit root. *J. Econ.* 136 (1), 115–130.
- Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. *Biometrika* 75 (2), 335–346.
- Phillips, P.C.B., Shi, S., 2017. Detecting financial collapse and ballooning sovereign risk. In: Cowles Foundation Discussion Paper No. 2110. Available at SSRN: <https://ssrn.com/abstract=3036545>.
- Phillips, P.C.B., Shi, S., 2018. Financial bubble implosion and reverse regression. *Econ. Theory* 34 (4), 705–753.
- Phillips, P.C.B., Solo, V., 1992. Asymptotics for linear processes. *Ann. Stat.* 20, 971–1001.
- Phillips, P.C.B., Yu, J., 2009. Limit Theory for Dating the Origination and Collapse of Mildly Explosive Periods in Time Series Data. Singapore Management University. Unpublished Manuscript.
- Phillips, P.C.B., Yu, J., 2011. Warning Signs of Future Asset Bubbles. The Straits Times, Singapore.
- Phillips, P.C.B., Yu, J., 2013. Bubble or roller coaster in world stock markets. The Business Times, Singapore.
- Phillips, P.C.B., Wu, Y., Yu, J., 2011. Explosive behavior in the 1990s Nasdaq: when did exuberance escalate asset values? *Int. Econ. Rev.* 52 (1), 201–226.
- Phillips, P.C.B., Shi, S., Yu, J., 2014. Specification sensitivity in right-tailed unit root testing for explosive behaviour. *Oxf. Bull. Econ. Stat.* 76 (3), 315–333.
- Phillips, P.C.B., Shi, S., Yu, J., 2015a. Testing for multiple bubbles: historical episodes of exuberance and collapse in the S&P 500. *Int. Econ. Rev.* 56 (4), 1043–1078.
- Phillips, P.C.B., Shi, S., Yu, J., 2015b. Testing for multiple bubbles: limit theory of real-time detectors. *Int. Econ. Rev.* 56 (4), 1079–1134.
- Phillips, P.C.B., Shi, S., Caspi, I., 2018. Real-Time Monitoring of Asset Markets with R. R Foundation for Statistical Computing, Vienna, Austria. URL <https://CRAN.R-project.org/package=psymonitor>.
- Shi, S., 2007. Moving Window Unit Root Test: Locating Real Estate Price Bubbles in Seoul Apartment Market. Master’s thesis, Singapore Management University, Singapore.
- Shi, S., 2017. Speculative bubbles or market fundamentals? an investigation of US regional housing markets. *Econ. Model.* 66, 101–111.
- Shi, S., Song, Y., 2016. Identifying speculative bubbles with an infinite hidden markov model. *J. Financ. Econom.* 14 (1), 159–184.
- Shi, S., Valadkhani, A., Smyth, R., Vahid, F., 2016. Dating the timeline of house price bubbles in Australian capital cities. *Econ. Rec.* 92 (299), 590–605.
- Shi, S., Hurn, S., Phillips, P.C.B., 2018. Causal Change Detection in Possibly Integrated Systems: Revisiting the Money-Income Relationship. Available at SSRN: <https://ssrn.com/abstract=3237213>.

- Vasilopoulos, K., Pavlidis, E., Spavound, S., 2018. Econometric Analysis of Explosive Time Series. R Foundation for Statistical Computing, Vienna, Austria. URL <https://github.com/kvasilopoulos/exuber>.
- Yiu, M.S., Jin, L., 2013. Detecting bubbles in the Hong Kong residential property market: an explosive-pattern approach. *J. Asian Econ.* 28 (1), 115–124.

Further reading

- Harvey, D.I., Leybourne, S.J., Zu, Y., 2018a. Sign-Based Unit Root Tests for Explosive Financial Bubbles in the Presence of Nonstationary Volatility. Working paper, University of Nottingham.

Chapter 3

Component-wise AdaBoost algorithms for high-dimensional binary classification and class probability prediction

Jianghao Chu*, Tae-Hwy Lee* and Aman Ullah*

Department of Economics, University of California, Riverside, CA, United States

*Corresponding authors: e-mail: jianghao.chu@email.ucr.edu; taelee@ucr.edu; aman.ullah@ucr.edu

Abstract

Freund and Schapire (1997) introduced “Discrete AdaBoost” (DAB) which has been mysteriously effective for the high-dimensional binary classification or binary prediction. In an effort to understand the myth, Friedman, Hastie, and Tibshirani (FHT, 2000) show that DAB can be understood as statistical learning which builds an additive logistic regression model via Newton-like updating minimization of the “exponential loss.” From this statistical point of view, FHT proposed three modifications of DAB, namely, Real AdaBoost (RAB), LogitBoost (LB), and Gentle AdaBoost (GAB). All of DAB, RAB, LB, GAB solve for the logistic regression via different algorithmic designs and different objective functions. The RAB algorithm uses class probability estimates to construct real-valued contributions of the weak learner, LB is an adaptive Newton algorithm by stagewise optimization of the Bernoulli likelihood, and GAB is an adaptive Newton algorithm via stagewise optimization of the exponential loss. The same authors of FHT published an influential textbook, *The Elements of Statistical Learning* (ESL, 2001 and 2008). A companion book *An Introduction to Statistical Learning* (ISL) by James et al. (2013) was published with applications in R. However, both ESL and ISL (e.g., sections 4.5 and 4.6) do not cover these four AdaBoost algorithms while FHT provided some simulation and empirical studies to compare these methods. Given numerous potential applications, we believe it would be useful to collect the R libraries of these AdaBoost algorithms, as well as more recently developed extensions to AdaBoost for probability prediction with examples and illustrations. Therefore, the goal of this chapter is to do just that, i.e., (i) to provide a user guide of these alternative AdaBoost algorithms with step-by-step tutorial of using R (in a way similar to ISL, e.g., section 4.6), (ii) to compare AdaBoost with alternative

machine learning classification tools such as the Deep Neural Network (DNN), logistic regression with LASSO and SIM-RODEO, and (iii) to demonstrate the empirical applications in economics, such as prediction of business cycle turning points and directional prediction of stock price indexes. We revisit Ng (2014) who used DAB for prediction of the business cycle turning points by comparing the results from RAB, LB, GAB, DNN, logistic regression, and SIM-RODEO.

Keywords: AdaBoost, R, Binary classification, Logistic regression, DAB, RAB, LB, GAB, DNN

1 Introduction

A large number of important variables in economics are binary. Let

$$\pi(x) \equiv P(y = 1|x)$$

and y takes value 1 with probability $\pi(x)$ and -1 with probability $1 - \pi(x)$. The studies on making the best forecast on y can be classified into two classes (Lahiri and Yang, 2012). One is focusing on getting the right probability model $\hat{\pi}(x)$, e.g., logit and probit models (Bliss, 1934; Cox, 1958; Walker and Duncan, 1967), then making the forecast on y with $\hat{\pi}(x) > 0.5$ using the estimated probability model. The other is to get the optimal forecast rule on y directly, e.g., the maximum score approach (Elliott and Lieli, 2013; Manski, 1975, 1985), without having to (correctly) model the probability $\hat{\pi}(x)$.

Given the availability of high-dimensional data, the binary classification or binary probability prediction problems can be improved by incorporating a large number of covariates (x). A number of new methods are proposed to take advantage of the great number of covariates. Freund and Schapire (1997) introduce machine learning method called Discrete AdaBoost algorithm, which takes a functional descent procedure and selects the covariates (or predictors) sequentially. Friedman et al. (2000) show that AdaBoost can be understood as a regularized logistic regression, which selects the covariates one-at-a-time. The influential paper also discusses several extensions to the original idea of Discrete AdaBoost and proposes new Boosting methods, namely Real AdaBoost, LogitBoost, and Gentle Boost, which uses the exponential loss or Bernoulli log-likelihood as fitting criteria. Later on, Friedman (2001) generalizes the idea to any fitting criteria and proposes the Gradient Boosting Machine. Bühlmann and Yu (2003) and Bühlmann (2006) propose the L_2 Boost and prove its consistency for regression and classification. Mease et al. (2007) use the logistic function to convert the class label output of boosting algorithms into probability and/or quantile predictions. Chu et al. (2018a) show the linkage between the Discrete AdaBoost and the maximum score approach and propose Asymmetric AdaBoost for utility based high-dimensional binary classification.

On the other hand, efforts are made to incorporate traditional binary classification and probability prediction methods into the high-dimensional sparse

matrix setup. The key feature of high-dimensional data is the redundancy of covariates in the data. Hence, methods are proposed to select useful covariates while/before estimation of the models. Tibshirani (1996) proposes the LASSO that is to add L_1 penalty to including more covariates in the model. Zou (2006) derives a necessary condition for consistency of the LASSO variable selection and proposes the Adaptive LASSO which is shown to have the oracle property. LASSO-type methods are often used with parametric models such as linear model or logistic model. To relax the parametric assumptions, Lafferty and Wasserman (2008) propose the Regularization of the Derivative Expectation Operator (RODEO) for variable selection in kernel regression. Chu et al. (2018b) propose SIM-RODEO for variable selection in semiparametric single-index model. See Su and Zhang (2014) for a thorough review of variable selection in nonparametric and semiparametric models. James et al. (2013) also give a comprehensive introduction to the literature.

This chapter gives an overview of recently developed machine learning methods, namely AdaBoost, in the role of binary prediction. AdaBoost algorithm focuses on making the optimal forecast directly without modeling the conditional probability of the events. AdaBoost gets an additive model by iteratively minimizing an exponential loss function. In each iteration, AdaBoost puts more weights on the observations that cannot be predicted correctly using the previous predictors. Moreover, AdaBoost algorithm is able to solve classification problem with high-dimensional data which is an advantage to traditional classification methods.

The rest of the chapter is organized as follows. In Section 2 we provide a brief introduction of AdaBoost from minimizing the “exponential loss.” In Section 3 we show popular variants of AdaBoost. Section 4 introduces alternative methods for AdaBoost. Section 5 gives numerical examples of the boosting algorithms. Section 6 compares the mentioned boosting algorithms with Deep Neural Network (DNN) and logistic regression with LASSO. Section 7 concludes.

2 AdaBoost

The algorithm of AdaBoost is shown in Algorithm 1. Let y be the binary class taking a value in $\{-1, 1\}$ that we wish to predict. Let $f_m(x)$ be the weak learner (weak classifier) for the binary target y that we fit to predict using the high-dimensional covariates x in the m th iteration. Let err_m denote the error rate of the weak learner $f_m(x)$, and $E_w(\cdot)$ denote the weighted expectation (to be defined below) of the variable in the parenthesis with weight w . Note that the error rate $E_w[1_{(y \neq f_m(x))}]$ is estimated by $err_m = \sum_{i=1}^n w_i 1_{(y_i \neq f_m(x_i))}$ with the weight w_i given by step 2(c) from the previous iteration. n is the number of observations. The symbol $1_{(\cdot)}$ is the indicator function which takes the

ALGORITHM 1 Discrete AdaBoost (DAB, Freund and Schapire, 1997)

1. Start with weights $w_i = \frac{1}{n}, i = 1, \dots, n$.
2. For $m = 1$ to M
 - (a) For $j = 1$ to k (for each variable)
 - i. Fit the classifier $f_{mj}(x_{ij}) \in \{-1, 1\}$ using weights w_i on the training data.
 - ii. Compute $err_{mj} = \sum_{i=1}^n w_i \mathbf{1}_{(y_i \neq f_{mj}(x_{ji}))}$.
 - (b) Find $\hat{j}_m = \arg \min_j err_{mj}$
 - (c) Compute $c_m = \log \left(\frac{1 - err_{m, \hat{j}_m}}{err_{m, \hat{j}_m}} \right)$.
 - (d) Set $w_i \leftarrow w_i \exp [c_m \mathbf{1}_{(y_i \neq f_{m, \hat{j}_m}(x_{j_{m,i}}))}], i = 1, \dots, n$, and normalize so that $\sum_{i=1}^n w_i = 1$.
3. Output the binary classifier $\text{sign}[F_M(x)]$ and the class probability prediction $\hat{\pi}(x) = \frac{e^{F_M(x)}}{e^{F_M(x)} + e^{-F_M(x)}}$ where $F_M(x) = \sum_{m=1}^M c_m f_{m, \hat{j}_m}(x_{j_m})$.

value 1 if a logical condition inside the parenthesis is satisfied and takes the value 0 otherwise. The symbol $\text{sign}(z) = 1$ if $z > 0$, $\text{sign}(z) = -1$ if $z < 0$, and hence $\text{sign}(z) = \mathbf{1}_{(z>0)} - \mathbf{1}_{(z<0)}$.

Remark 1. Note that the presented version of Discrete AdaBoost as well as Real AdaBoost (RAB), LogitBoost (LB), and Gentle AdaBoost (GAB) which will be introduced later in the chapter are different from their original version when they are first introduced. The original version of these algorithms only output the class label. In this chapter, we follow the idea of Mease et al. (2007) and modified the algorithms to output both the class label and probability prediction. The probability prediction is attained using

$$\hat{\pi}(x) = \frac{e^{F_M(x)}}{e^{F_M(x)} + e^{-F_M(x)}}, \quad (1)$$

where $F_M(x)$ is the sum of weak learner in the algorithms.

The most widely used weak learner is the classification tree. The simplest classification tree, the stump, takes the following functional form

$$f(x_j, a) = \begin{cases} 1 & x_j > a \\ -1 & x_j \leq a \end{cases}$$

where the parameter a is found by minimizing the error rate

$$\min_a \sum_{i=1}^n w_i \mathbf{1}_{(y_i \neq f(x_{ji}, a))}. \quad (2)$$

In addition to the commonly used classification tree weak learners in machine learning literature described above, Discrete AdaBoost, in principle, can take any classifier and boost its performance through the weighted voting scheme. For example, we can also use a one-variable Logistic Regression as a weak learner which we will call the logistic weak learner. Simulation results of [Chu et al. \(2018a\)](#) show that the logistic weak learner generally has better performance than the stump in traditional econometric models. In the logistic weak learner, we assume the probability

$$\pi(x_j) \equiv P(y = 1|x_j) = \frac{e^{x_j\beta}}{1 + e^{x_j\beta}}.$$

Let $Y = \frac{y+1}{2} \in \{0, 1\}$. We estimate the parameter β by maximizing the weighted logistic log-likelihood function

$$\begin{aligned} \max_{\beta} \log L &= \log \prod_{i=1}^n \left[\left(\frac{e^{x_{ji}\beta}}{1 + e^{x_{ji}\beta}} \right)^{Y_i} \left(\frac{1}{1 + e^{x_{ji}\beta}} \right)^{1-Y_i} \right]^{w_i} \\ &= \log \prod_{i=1}^n \left(\frac{e^{Y_i x_{ji}\beta}}{1 + e^{x_{ji}\beta}} \right)^{w_i} \end{aligned} \quad (3)$$

$$\begin{aligned} &= \sum_{i=1}^n \log \left(\frac{e^{Y_i x_{ji}\beta}}{1 + e^{x_{ji}\beta}} \right)^{w_i} \\ &= \sum_{i=1}^n w_i [Y_i x_{ji}\beta - \log(1 + e^{x_{ji}\beta})]. \end{aligned} \quad (4)$$

Then the resulting logistic weak learner will be

$$f(x_j, \beta) = \begin{cases} 1 & \pi(x_j, \beta) > 0.5 \\ -1 & \pi(x_j, \beta) < 0.5. \end{cases}$$

Several packages in R provide off-the-shelf implementations of Discrete AdaBoost. *JOUSBoost* gives an implementation of the Discrete AdaBoost algorithm from [Freund and Schapire \(1997\)](#) applied to decision tree classifiers and provides a convenient function to generate test sample of the algorithm ([Olson, 2017](#)).

Here we use the *circle_data* function from *JOUSBoost* to generate a test sample. The *circle_data* function simulate draws from a Bernoulli distribution over $\{-1, 1\}$. First, the predictors x are drawn i.i.d. uniformly over the square in the two-dimensional plane centered at the origin with side length $2 * outer_r$, and then the response is drawn according to $\pi(x)$, which depends on $r(x)$, the euclidean norm of x . If $r(x) \leq inner_r$, then $\pi(x) = 1$, if $r(x) \geq outer_r$ then $\pi(x) = 0$, and $\pi(x) = (outer_r - r(x))/(outer_r - inner_r)$ when $inner_r \leq r(x) \leq outer_r$ as in [Mease et al. \(2007\)](#). The code of the function is shown below.

```
circle_data <- function (n = 500, inner_r = 8, outer_r = 28) {  
  if (outer_r <= inner_r)  
    stop("outer_r must be strictly larger than inner_r")  
  X = matrix(stats::runif(2 * n, -outer_r, outer_r),  
             nrow = n, ncol = 2)  
  r = apply(X, 1, function(x) sqrt(sum(x^2)))  
  p = 1 * (r < inner_r) + (outer_r - r)/(outer_r -  
    inner_r) *  
    ((inner_r < r) & (r < outer_r))  
  y = 2 * stats::rbinom(n, 1, p) - 1  
  list(X = X, y = y, p = p)  
}
```

Then we use the implementation of Discrete AdaBoost from *ada* package since the *ada* package provides implementation of not only Discrete AdaBoost and also Real AdaBoost, LogitBoost, and Gentle AdaBoost which we will discuss about in the next section ([Culp et al., 2016](#)).

```
#Generate data from the circle model  
library(JOUSBoost)  
set.seed(111)  
dat <- circle_data(n = 500)  
x <- dat$X  
y <- dat$y  
  
library(ada)  
model <- ada(x, y, loss = "exponential", type = "discrete",  
iter = 200)  
print(model)
```

where *y* and *x* are the training samples, and *iter* controls the number of boosting iterations.

Remark 2. The algorithms in *ada* for Discrete AdaBoost, Real AdaBoost, LogitBoost, and Gentle Boost may not follow exactly the same steps and/or criteria as described in the chapter. However, the major settings, the loss function, and characteristics of weak learners, are the same. We choose to use the *ada* package since it is widely accessible and easy to use for the readers.

The output is as follows.

```
Call:  
ada(x, y = y, loss = "exponential", type = "discrete", iter = 200)  
Loss: exponential Method: discrete Iteration: 200  
Final Confusion Matrix for Data:
```

```

    Final Prediction
True value   -1      1
      -1     300    14
      1      15    171

Train Error: 0.058

Out-Of-Bag Error: 0.094 iteration= 195

Additional Estimates of number of iterations:

train.err1 train.kap1
 197         197

```

Other packages include *fastAdaboost* (Chatterjee, 2016) which uses C++ code in the backend to provide an implementation of AdaBoost that is about 100 times faster than native R based libraries.

```

library(fastAdaboost)
adaboost(y~x, nIter)

```

where y and x are the training samples and $nIter$ is the number of boosting iterations. Note that *fastAdaboost* also contains implementation of Real AdaBoost which we will introduce later. *GBM* which is short for Generalized Boosting Regression Models contains implementation of extensions to Freund and Schapire's AdaBoost algorithm and Friedman's gradient boosting machine (Ridgeway, 2017).

Friedman et al. (2000) show that AdaBoost builds an additive logistic regression model

$$F_M(x) = \sum_{m=1}^M c_m f_m(x) \quad (5)$$

via Newton-like updates for minimizing the exponential loss

$$J(F) = E(e^{-yF(x)}|x). \quad (6)$$

We use a greedy method to minimize the exponential loss function iteratively. After m iterations, the current classifier is denoted as $F_m(x) = \sum_{s=1}^m c_s f_s(x)$. In the next iteration, we are seeking an update $c_{m+1} f_{m+1}(x)$ for the function fitted from previous iterations $F_m(x)$. The updated classifier would take the form

$$F_{m+1}(x) = F_m(x) + c_{m+1} f_{m+1}(x).$$

The loss for $F_{m+1}(x)$ will be

$$\begin{aligned} J(F_{m+1}(x)) &= J(F_m(x) + c_{m+1} f_{m+1}(x)) \\ &= E\left[e^{-y(F_m(x) + c_{m+1} f_{m+1}(x))}\right]. \end{aligned} \quad (7)$$

Expand w.r.t. $f_{m+1}(x)$

$$\begin{aligned} J(F_{m+1}(x)) &\approx E \left[e^{-yF_m(x)} \left[1 - yc_{m+1} f_{m+1}(x) + \frac{y^2 c_{m+1}^2 f_{m+1}^2(x)}{2} \right] \right] \\ &= E \left[e^{-yF_m(x)} \left(1 - yc_{m+1} f_{m+1}(x) + \frac{c_{m+1}^2}{2} \right) \right]. \end{aligned}$$

The last equality holds since $y \in \{-1, 1\}$, $f_{m+1}(x) \in \{-1, 1\}$, and $y^2 = f_{m+1}^2(x) = 1$. $f_{m+1}(x)$ only appears in the second term in the parenthesis, so minimizing the loss function (7) w.r.t. $f_{m+1}(x)$ is equivalent to maximizing the second term in the parenthesis which results in the following conditional expectation

$$\max_f E \left[e^{-yF_m(x)} yc_{m+1} f_{m+1}(x) | x \right].$$

For any $c > 0$ (we will prove this later), we can omit c_{m+1} in the above objective function

$$\max_f E \left[e^{-yF_m(x)} y f_{m+1}(x) | x \right].$$

To compare it with the Discrete AdaBoost algorithm, here we define weight $w = w(y, x) = e^{-yF_m(x)}$. Later we will see that this weight w is equivalent to that shown in the Discrete AdaBoost algorithm. So the above optimization can be seen as maximizing a weighted conditional expectation

$$\max_f E_w[y f_{m+1}(x) | x] \tag{8}$$

where $E_w(y|x) := \frac{E(wy|x)}{E(w|x)}$ refers to a weighted conditional expectation. Note that (8)

$$\begin{aligned} E_w[y f_{m+1}(x) | x] &= P_w(y=1|x) f_{m+1}(x) - P_w(y=-1|x) f_{m+1}(x) \\ &= [P_w(y=1|x) - P_w(y=-1|x)] f_{m+1}(x) \\ &= E_w(y|x) f_{m+1}(x). \end{aligned}$$

where $P_w(y|x) = \frac{E(w|y,x)P(y|x)}{E(w|x)}$. Solve the maximization problem (8). Since $f_{m+1}(x)$ only takes 1 or -1, it should be positive whenever $E_w(y|x)$ is positive and -1 whenever $E_w(y|x)$ is negative. The solution for $f_{m+1}(x)$ is

$$f_{m+1}(x) = \begin{cases} 1 & E_w(y|x) > 0 \\ -1 & \text{otherwise.} \end{cases}$$

Next, minimize the loss function (7) w.r.t. c_{m+1}

$$c_{m+1} = \arg \min_{c_{m+1}} E_w\left(e^{-c_{m+1}yf_{m+1}(x)}\right)$$

$$E_w\left(e^{-c_{m+1}yf_{m+1}(x)}\right) = P_w(y=f_{m+1}(x))e^{-c_{m+1}} + P_w(y \neq f_{m+1}(x))e^{c_{m+1}}$$

$$\frac{\partial E_w\left(e^{-c_{m+1}yf_{m+1}(x)}\right)}{\partial c} = -P_w(y=f_{m+1}(x))c_{m+1}e^{-c_{m+1}} + P_w(y \neq f_{m+1}(x))c_{m+1}e^{c_{m+1}}$$

Let

$$\frac{\partial E_w\left(e^{-c_{m+1}yf_{m+1}(x)}\right)}{\partial c_{m+1}} = 0,$$

and we have

$$P_w(y=f_{m+1}(x))c_{m+1}e^{-c_{m+1}} = P_w(y \neq f_{m+1}(x))c_{m+1}e^{c_{m+1}},$$

Solve for c_{m+1} , we obtain

$$c_{m+1} = \frac{1}{2} \log \frac{P_w(y=f_{m+1}(x))}{P_w(y \neq f_{m+1}(x))} = \frac{1}{2} \log \left(\frac{1 - err_{m+1}}{err_{m+1}} \right),$$

where $err_{m+1} = P_w(y \neq f_{m+1}(x))$ is the error rate of $f_{m+1}(x)$. Note that $c_{m+1} > 0$ as long as the error rate is smaller than 50%. Our assumption $c_{m+1} > 0$ holds for any learner that is better than random guessing.

Now we have finished the steps of one iteration and can get our updated classifier by

$$F_{m+1}(x) \leftarrow F_m(x) + \left(\frac{1}{2} \log \left(\frac{1 - err_{m+1}}{err_{m+1}} \right) \right) f_{m+1}(x).$$

Note that in the next iteration, the weight we defined w_{m+1} will be

$$w_{m+1} = e^{-yF_{m+1}(x)} = e^{-y(F_m(x) + c_{m+1}f_{m+1}(x))} = w_m \times e^{-c_{m+1}f_{m+1}(x)y}.$$

Since $-yf_{m+1}(x) = 2 \times 1_{\{y \neq f_{m+1}(x)\}} - 1$, the update is equivalent to

$$w_{m+1} = w_m \times e^{\left(\log \left(\frac{1 - err_{m+1}}{err_{m+1}} \right) 1_{[y \neq f_{m+1}(x)]} \right)} = w_m \times \left(\frac{1 - err_{m+1}}{err_{m+1}} \right)^{1_{[y \neq f_{m+1}(x)]}}.$$

Thus the function and weights update are of an identical form to those used in AdaBoost. AdaBoost could do better than any single weak classifier since it iteratively minimizes the loss function via a Newton-like procedure. Interestingly, the function $F(x)$ from minimizing the exponential loss is the same as maximizing a logistic log-likelihood. Let

$$J(F(x)) = E\left[E\left(e^{-yF(x)}|x\right)\right]$$

$$= E\left[P(y=1|x)e^{-F(x)} + P(y=-1|x)e^{F(x)}\right].$$

Taking derivative w.r.t. $F(x)$ and making it equal to zero, we obtain

$$\begin{aligned}\frac{\partial E(e^{-yF(x)}|x)}{\partial F(x)} &= -P(y=1|x)e^{-F(x)} + P(y=-1|x)e^{F(x)} = 0 \\ F^*(x) &= \frac{1}{2} \log \left[\frac{P(y=1|x)}{P(y=-1|x)} \right].\end{aligned}$$

Moreover, if the true probability

$$P(y=1|x) = \frac{e^{2F(x)}}{1+e^{2F(x)}},$$

for $Y = \frac{y+1}{2}$, the log-likelihood is

$$E(\log L|x) = E[2YF(x) - \log(1+e^{2F(x)})|x].$$

The solution $F^*(x)$ that maximize the log-likelihood must equals the $F(x)$ in the true model $P(y=1|x) = \frac{e^{2F(x)}}{1+e^{2F(x)}}$. Hence,

$$\begin{aligned}e^{2F^*(x)} &= P(y=1|x) \left(1 + e^{2F^*(x)} \right) \\ e^{2F^*(x)} &= \frac{P(y=1|x)}{1 - P(y=1|x)} \\ F^*(x) &= \frac{1}{2} \log \left[\frac{P(y=1|x)}{P(y=-1|x)} \right].\end{aligned}\tag{9}$$

AdaBoost that minimizes the exponential loss yield the same solution as logistic regression that maximizes the logistic log-likelihood.

3 Extensions to AdaBoost algorithms

In this section we introduce extensions of Discrete AdaBoost, namely Real AdaBoost (RAB), LogitBoost (LB), and Gentle AdaBoost (GAB) and discuss how some aspects of the DAB may be modified to yield RAB, LB, and GAB. In the last section, we learned that Discrete AdaBoost minimizes an exponential loss via iteratively adding a binary weaker learner to the pool of weak learners. The addition of a new weak learner can be seen as taking a step on the direction that loss function descents in the Newton method. There are two major ways to extend the idea of Discrete AdaBoost. One focuses on making the minimization method more efficient by adding a more flexible weak learner. The other is to use different loss functions that may lead to better results. Next, we give an introduction to several extensions of Discrete AdaBoost.

3.1 Real AdaBoost

ALGORITHM 2 Real AdaBoost (RAB, Friedman et al., 2000)

1. Start with weights $w_i = \frac{1}{n}, i = 1, \dots, n$.
2. For $m = 1$ to M
 - (a) For $j = 1$ to k (for each variable)
 - i. Fit the classifier to obtain a class probability estimate $p_m(x_j) = \hat{P}_w(y=1|x_j) \in [0, 1]$ using weights w_i on the training data.
 - ii. Let $f_{mj}(x_j) = \frac{1}{2} \log \frac{p_m(x_j)}{1-p_m(x_j)}$.
 - iii. Compute $err_{mj} = \sum_{i=1}^n w_i \mathbf{1}_{(y_i \neq \text{sign}(f_{mj}(x_j)))}$.
 - (b) Find $\hat{j}_m = \arg \min_j err_{mj}$.
 - (c) Set $w_i \leftarrow w_i \exp[-y_i f_{m, \hat{j}_m}(x_{\hat{j}_m}, i)], i = 1, \dots, n$, and normalize so that $\sum_{i=1}^n w_i = 1$.
3. Output the classifier $\text{sign}[F_M(x)]$ and the class probability prediction $\hat{\pi}(x) = \frac{e^{F_M(x)}}{e^{F_M(x)} + e^{-F_M(x)}}$ where $F_M(x) = \sum_{m=1}^M f_m(x)$.

Real AdaBoost focuses solely on improving the minimization procedure of Discrete AdaBoost. In Real AdaBoost, the weak learners are continuous comparing to Discrete AdaBoost where the weak learners are binary (discrete). Real AdaBoost is minimizing the exponential loss with continuous updates where Discrete AdaBoost minimizes the exponential loss with discrete updates. Hence, Real AdaBoost is more flexible with the step size and direction of the minimization and minimizes the exponential loss faster and more accurately. However, Real AdaBoost also imposes restriction that the classifier must produce a probability prediction which reduces the flexibility of the model. As we shall see in the numerical examples, Real AdaBoost may achieve a larger in-sample training error due to the flexibility of its model. On the other hand, this also reduces the chances of fitting and would in the end achieve a smaller out-of-sample test error.

As we mentioned earlier, *ada* gives an implementation of the Real AdaBoost as well as Discrete AdaBoost.

```
#Generate data from the circle model
library(JOUSBoost)
set.seed(111)
dat <- circle_data(n = 500)
x <- dat$x
y <- dat$y

library(ada)
model <- ada(x, y, loss = "exponential", type = "real", iter = 200)
print(model)
```

where y and x are the training samples, and $iter$ controls the number of boosting iterations. The output is as follows.

```

Call:
ada(x, y = y, loss = "exponential", type = "real", iter = 200)

Loss: exponential Method: real Iteration: 200

Final Confusion Matrix for Data:
      Final Prediction

True value   -1     1
      -1    293    21
      1     29   157

Train Error: 0.1

Out-Of-Bag Error: 0.114 iteration= 189

Additional Estimates of number of iterations:

train.err1 train.kap1
189         189

```

3.2 LogitBoost

Friedman et al. (2000) propose LogitBoost by minimizing the Bernoulli log-likelihood via an adaptive Newton algorithm for fitting an additive logistic regression model. LogitBoost extends Discrete AdaBoost in two ways. First, it uses the Bernoulli log-likelihood instead of exponential function as loss function. Furthermore, it updates the classifier by adding a linear model instead of a binary weak learner.

ALGORITHM 3 LogitBoost (LB, Friedman et al., 2000)

1. Start with weights $w_i = \frac{1}{n}, i = 1, \dots, n$, $F(x) = 0$ and probability estimates $p(x_i) = \frac{1}{2}$.
2. For $m = 1$ to M
 - (a) Compute the working response and weights
$$z_i = \frac{y_i^* - p(x_i)}{p(x_i)(1 - p(x_i))} \quad (10)$$

$$w_i = p(x_i)(1 - p(x_i)) \quad (11)$$
 - (b) For $j = 1$ to k (for each variable)
 - i. Fit the function $f_{mj}(x_j)$ by a weighted least-squares regression of z_i to x_j using weights w_i on the training data.
 - ii. Compute $err_{mj} = 1 - R^2$ from the weighted least-squares regression.
 - (c) Find $\hat{j}_m = \arg \min_j err_{mj}$
 - (d) Update $F(x) \leftarrow F(x) + \frac{1}{2} f_{m,\hat{j}_m}(x_{\hat{j}_m})$ and $p(x) \leftarrow \frac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}}$, $i = 1, \dots, n$.
3. Output the classifier $\text{sign}[F_M(x)]$ and the class probability prediction $\hat{\pi}(x) = \frac{e^{F_M(x)}}{e^{F_M(x)} + e^{-F_M(x)}}$ where $F_M(x) = \sum_{m=1}^M f_{m,\hat{j}_m}(x_{\hat{j}_m})$.

In LogitBoost, continuous weak learner is used similar to Real AdaBoost. However, LogitBoost specified the use of linear weak learner while Real AdaBoost allows any weak learner that returns a probability between zero and one. A bigger and more fundamental difference here is that LogitBoost uses the Bernoulli log-likelihood as loss function instead of the exponential loss. Hence, LogitBoost is more similar to logistic regression than Discrete AdaBoost and Real AdaBoost. As we will see in the simulation result, LogitBoost has the smallest in-sample training error but the largest out-of-sample test error. This implies that while LogitBoost is the most flexible of the four, it suffers the most from overfitting.

LogitBoost is arguably one of the most well-known boosting algorithm. Popular packages are available such as *caTools* (Tuszynski, 2018). For consistency of the chapter, here we stick with the *ada* package which gives an ideal implementation of LogitBoost algorithm for small to moderate-sized data sets.

```
#Generate data from the circle model
library(JOUSBoost)
set.seed(111)
dat <- circle_data(n = 500)
x <- dat$X
y <- dat$y

library(ada)
model <- ada(x, y, loss = "logistic", type = "gentle", iter = 200)
print(model)
```

where y and x are the training samples, and $iter$ controls the number of boosting iterations. The output is as follows.

```
Call:
ada(x, y = y, loss = "logistic", type = "gentle", iter = 200)

Loss: logistic Method: gentle Iteration: 200

Final Confusion Matrix for Data:
          Final Prediction
True value      -1      1
      -1     309     5
      1      8    178

Train Error: 0.026
Out-Of-Bag Error: 0.07 iteration= 196

Additional Estimates of number of iterations:

train.err1 train.kap1
195           195
```

3.3 Gentle AdaBoost

ALGORITHM 4 Gentle AdaBoost (GAB, Friedman et al., 2010)

1. Start with weights $w_i = \frac{1}{n}, i = 1, \dots, n$.
2. For $m = 1$ to M
 - (a) For $j = 1$ to k (for each variable)
 - i. Fit the regression function $f_{mj}(x_{ij})$ by weighted least-squares of y_i on x_i using weights w_i on the training data.
 - ii. Compute $err_{mj} = 1 - R^2$ from the weighted least-squares regression.
 - (b) Find $\hat{j}_m = \arg \min_j err_{mj}$
 - (c) Set $w_i \leftarrow w_i \exp[-y_i f_{m, \hat{j}_m}(x_{\hat{j}_m, i})], i = 1, \dots, n$, and normalize so that $\sum_{i=1}^n w_i = 1$.
3. Output the classifier $\text{sign}[F_M(x)]$ and the class probability prediction $\hat{\pi}(x) = \frac{e^{F_M(x)}}{e^{F_M(x)} + e^{-F_M(x)}}$ where $F_M(x) = \sum_{m=1}^M f_{m, \hat{j}_m}(x_{\hat{j}_m})$.

Gentle AdaBoost extends Discrete AdaBoost in the sense that it allows each weak learner to be a linear model. This is similar to LogitBoost and more flexible than Discrete AdaBoost and Real AdaBoost. However, it is closer to Discrete AdaBoost and Real AdaBoost than LogitBoost in the sense that Gentle AdaBoost, Discrete AdaBoost, and Real AdaBoost all minimize the exponential loss while LogitBoost minimizes the Bernoulli log-likelihood. Another point that Gentle AdaBoost is more similar to Real AdaBoost than Discrete AdaBoost is that since the weak learners are continuous, there is no need to find an optimal step size for each iteration because the weak learner is already optimal. As we will see in the simulation results, Gentle Boost often lies between Real AdaBoost and LogitBoost in terms of in-sample training error and out-of-sample test error.

ada also gives an implementation of the Gentle AdaBoost algorithm.

```
#Generate data from the circle model
library(JOUSBoost)
set.seed(111)
dat <- circle_data(n = 500)
x <- dat$X
y <- dat$y

library(ada)
model <- ada(x, y, loss = "exponential", type = "gentle",
iter = 200)
print(model)
```

where y and x are the training samples, and $iter$ controls the number of boosting iterations. The output is as follows.

```

Call:
ada(x, y = y, loss = "exponential", type = "gentle", iter = 200)

Loss: exponential Method: gentle Iteration: 200

Final Confusion Matrix for Data:
Final Prediction
True value    -1     1
-1      305     9
1       15   171

Train Error: 0.048

Out-Of-Bag Error: 0.078 iteration= 198

Additional Estimates of number of iterations:

train.err1 train.kap1
196          196

```

For all the four boosting algorithms mentioned above, *ada* outputs the class label by default. However, we can use the command *predict* to output probability prediction and/or of class label using *ada*.

```

#Generate data from the circle model
library(JOOSBoost)
set.seed(111)
dat <- circle_data(n = 500)
x <- dat$x
y <- dat$y

library(ada)
model <- ada(x, y, loss = "exponential", type = "discrete",
iter = 200)

#New Data for Prediction
newx <- data.frame(1,1)
names(newx) <- c('V1', 'V2')
predict(model, newdata = newx, type = "F")
predict(model, newdata = newx, type = "prob")

```

where *y* and *x* are the training samples, *iter* controls the number of boosting iterations. *model* is the output from fitting the model using *ada*, *newdata* is the data to be used in prediction and *type* specifies the type of output from the *predict* function. When *type* = “vector”, the function outputs class labels. When *type* = “*F*”, the function outputs $F(x)$ which the sum of all weak learners. When *type* = “prob”, the function outputs the class probability using (1). The output is as follows.

```
> predict(model, newdata = newx, type = "F")
1
4.345358
> predict(model, newdata = newx, type = "prob")
[,1]      [,2]
1 0.0001681114 0.9998319
```

Note that manually transforming the sum of all weak learners $F(x)$ into probability prediction using Eq. (9) would lead to the same result as directly output the probability prediction from the package as in the second line.

4 Alternative classification methods

Apart from Boosting algorithms, we also consider Deep Neural Network, Logistic Regression, and semiparametric single-index model as alternative methods to obtain a predictor of y . Deep Neural Network is able to deal with high-dimensional data. For Logistic Regression, we have to select useful information from noises. Hence, a shrinkage parameter is used with the logistic log-likelihood which we call LASSO. Semiparametric single-index model is an extension to parametric single-index model such as Logistic Regression. It relaxes the parametric assumptions and uses the kernel function of fit the data locally. For high-dimensional problem, we use SIM-RODEO to select useful explanation variables for semiparametric single-index models.

4.1 Deep Neural Network

Deep Neural Network is undoubtedly one of the most state-of-the-art classification methods. The model is similar to a multistage regression or classification model. The idea is to build a flexible nonlinear statistical model consisted of several layers and each layer is consisted of neurons as in Fig. 1.

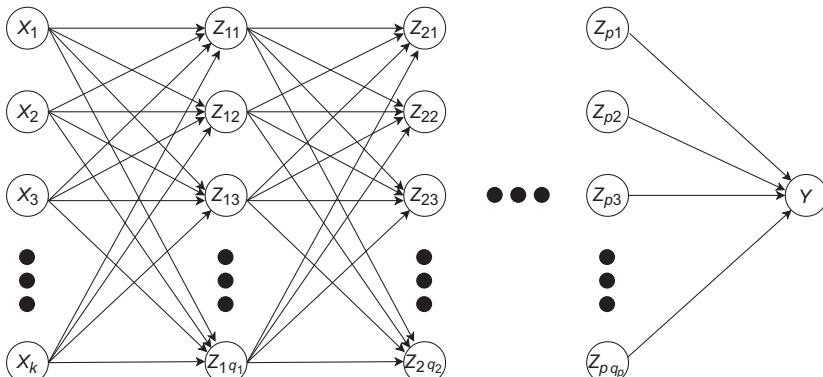


FIG. 1 Diagram of Deep Neural Network.

For binary classification, there is only one output Y that is the class probability or class label. Since the transformation from class probability to class label is straightforward, we focus on the case where the output is the class probability. The layer labeled X is the input layer which contains all the explanatory variables in the data set. Note that the number of explanatory variables k is allowed to be extremely large (larger than the number of observations) as in high-dimensional settings. The layers labeled Z are the hidden layers. The number of hidden layers p can be arbitrarily set by the user and each hidden layer can contain arbitrarily many neurons denoted by q_t where t stands for the t th hidden layer.

The output z_{ts} of the s th neuron in the t th hidden layer is normally a single-index function $g(\alpha_{ts} + \beta'_{ts} Z_{t-1})$ where α is a scalar, β is a vector of same length q_{t-1} as the number of neurons in the $(t-1)$ th hidden layer or the input layer if $t = 1$ and $Z_{t-1} = (z_{t-1,1}, z_{t-1,2}, \dots, z_{t-1,q_{t-1}})$ is a vector of outputs from all neurons of the $(t-1)$ th hidden layer or the input layer if $t = 1$. Similarly, the output layer of the model is also chosen to be a single-index function of the outputs of the last hidden layer. Hence,

$$z_{1s} = g(w_{01s} + w'_{1s} X) \quad (12)$$

$$z_{ts} = g(w_{0ts} + w'_{ts} Z_{t-1}) \quad (13)$$

$$Y = \hat{\pi}(x) = f(w_0 + w' Z_t). \quad (14)$$

The function $g(v)$ is called the activation function. It is often chosen to be a sigmoid. Popular choices are the Rectified Linear Unit (ReLU)

$$g(v) = \max(0, v)$$

and the logistic function

$$g(v) = \frac{1}{1 + e^{-v}}.$$

The function $f(v)$ in the output layer can also be a sigmoid. In addition to the ReLU and logistic function, the identity function can also be used as the output function.

Since the activation function, output function, and number of hidden layers and neurons are all chosen by the user prior to fitting the model, the only parameters to be determined by the data are the weights α 's and β 's. We choose the best values for α 's and β 's to minimize a given loss function. For binary classification, the squared error loss

$$L(w) = \sum_i (y_i - \hat{\pi}(x_i))^2$$

and the cross-entropy

$$L(w) = - \sum_i y_i \log \hat{\pi}(x_i)$$

are often used. The minimization procedure of Deep Neural Network is often time-consuming. Moreover, convergence and optimality can not be guaranteed. Hence, multiple attempts need to be made for a single problem. Two techniques, stochastic gradient descent and back-propagation, are often used for minimization of Deep Neural Network. Fortunately, we do not have to worry about the implementation of the minimization procedure since packages are available in R.

Remark 3. Note that the class probability can be converted to class label easily by the rule $\hat{Y} = 1(\hat{\pi}(x) > 0.5)$ where $1(\cdot)$ is the indicator function.

We now turn to the implementation of Deep Neural Network using R. There are two packages in R for Deep Neural Networks, *neuralnet* and *keras*. *neuralnet* is a package in R that solves Deep Neural Network (Fritsch and Guenther, 2016). *keras*, on the other hand, is an interface of tensorflow which we will introduce later in R. Hence, *neuralnet* is easier to use for R users and works fairly well on moderate-size problems. Let us introduce the use of *neuralnet* first.

```
#Generate data from the circle model
library(JOUSBoost)
set.seed(111)
dat <- circle_data(n = 500)
x <- dat$X
y <- dat$y > 0
sum.data <- data.frame(x,y)

library(neuralnet)
print(net.sum <- neuralnet(y~X1 + X2, sum.data, hidden = 2,
act.fct = "logistic", err.fct = "sse"))
```

where y is the class label, $X1$ and $X2$ are the explanatory variables, *hidden* is a vector that specifies the number of neurons in each layer and *act.fct* specifies the kind of activation function to be used. In our example, there is only one hidden layer with two neurons in it. The activation function is logistic and the loss function to be minimized is the sum of squared errors.

```
$call
neuralnet(formula = y ~ X1 + X2, data = sum.data, hidden = 2,
err.fct = "sse", act.fct = "logistic")

$response
y
1    TRUE
2   FALSE
3    TRUE
4   FALSE
5    TRUE
...
...
```

```

$covariate
[,1]      [,2]
[1,] 5.2069519311 4.2558353692
[2,] 12.6829428039 -20.3719270937
[3,] -7.2563678008 24.6251029056
[4,] 0.8357344829 -18.1280552782
[5,] -6.8508599121 17.9034926556
...
$model.list
$model.list$response
[1] "y"

$model.list$variables
[1] "X1" "X2"

$err.fct
function (x, y)
{
  1/2 * (y - x)^2
}
<bytecode: 0x4f10880>
<environment: 0x85c0258>
attr(,"type")
[1] "sse"

$act.fct
function (x)
{
  1/(1 + exp(-x))
}
<bytecode: 0x6b3fdf8>
<environment: 0x85c0258>
attr(,"type")
[1] "logistic"

$linear.output
[1] TRUE

$data
X1          X2      y
1 5.2069519311 4.2558353692 TRUE
2 12.6829428039 -20.3719270937 FALSE
3 -7.2563678008 24.6251029056 TRUE
4 0.8357344829 -18.1280552782 FALSE
5 -6.8508599121 17.9034926556 TRUE
...
$net.result
$net.result[[1]]

```

```

[,1]
1   0.580442461921
2   0.166226565353
3   0.746461567986
4   0.504631826509
5   0.880228917245
...
$weights
$weights[[1]]
$weights[[1]][[1]]
[,1]      [,2]
[1,]  0.9066077391 19.3160782267
[2,] -0.1145926600  1.3507916927
[3,]  0.0356239265 -0.3322296486

$weights[[1]][[2]]
[,1]
[1,] -0.8702072766
[2,]  1.0498484362
[3,]  0.8066945796

$startweights
$startweights[[1]]
$startweights[[1]][[1]]
[,1]      [,2]
[1,] -3.3233349646 -0.6039894538
[2,] -0.4675154531  0.6744466927
[3,]  0.4315402657  0.6359205358

$startweights[[1]][[2]]
[,1]
[1,] -0.6129703876
[2,]  0.4148913454
[3,]  0.8773433726

$generalized.weights
$generalized.weights[[1]]
[,1]      [,2]
1 -0.117151263217  0.036419330807
2 -0.148384521998  0.046128951957
3  0.910325193217 -0.221275093627
4 -0.119499601088  0.037149368973
5  0.071284648051 -0.011543348159
...
$result.matrix
1
error                      8.553736459613
reached.threshold           0.009821180033

```

```

steps           2145.000000000000
Intercept.to.1layhid1 0.906607739056
X1.to.1layhid1 -0.114592659965
X2.to.1layhid1 0.035623926505
Intercept.to.1layhid2 19.316078226658
X1.to.1layhid2 1.350791692688
X2.to.1layhid2 -0.332229648639
Intercept.to.y -0.870207276552
1layhid.1.to.y 1.049848436176
1layhid.2.to.y 0.806694579605

attr(,"class")
[1] "nn"

```

where ... represents that the rest of the output for this feature is omitted. The *keras* package (Allaire and Chollet, 2018) is a high-level interface of tensorflow which is an open source machine learning framework maintained by Google and is the most used library for fitting Deep Neural Networks. *keras* defines the structure and features of the neural network and send the informations to tensorflow which then solves the minimization problem and returns the results. *keras* is suitable for more complicated and larger problems since tensorflow actually doing the hard work.

Now we give a simple illustration of constructing neural networks with *keras* by constructing the same network as in the previous example in *keras*. More details about *keras* can be found at <https://tensorflow.rstudio.com>.

```

#Generate data from the circle model
library(JOUSBoost)
set.seed(111)
dat <- circle_data(n = 200)
x <- dat$X
y <- dat$y > 0
sum.data <- data.frame(x,y)

library(keras)
x_train <- x[1:100,]
y_train <- y[1:100,]
x_test <- x[101:200,]
y_test <- y[101:200,]

model <- keras_model_sequential()
model %>%
layer_dense(units = 2, activation = 'sigmoid', input_shape =
c(100)) %>%
layer_dense(units = 1, activation = 'softmax')
model %>% compile(
loss = 'categorical_crossentropy',

```

```

optimizer = optimizer_rmsprop(),
metrics = c('accuracy')
)

history <- model %>% fit(
x_train, y_train,
epochs = 30, batch_size = 100,
validation_split = 0.2
)

```

4.2 Logistic regression with LASSO

In traditional econometrics, the most used classification and probability prediction method should be logistic regression. Logistic regression assumes that the probability that the output variable $Y = \frac{y+1}{2} \in \{0,1\}$ takes value one follows a logistic function of x . That is

$$\pi(x) = P(Y = 1|x) = \frac{1}{1 + e^{-x\beta}}.$$

Given a sample data of y and x , the likelihood of the sample can be rewritten as

$$L(\beta) = \prod_i \left(\frac{1}{1 + e^{-x_i\beta}} \right)^{Y_i} \left(\frac{1}{1 + e^{x_i\beta}} \right)^{1-Y_i}. \quad (15)$$

Taking the log transformation, the log-likelihood is

$$\log L(\beta) = \log \left(\prod_i \left(\frac{1}{1 + e^{-x_i\beta}} \right)^{Y_i} \left(\frac{1}{1 + e^{x_i\beta}} \right)^{1-Y_i} \right) \quad (16)$$

$$= \sum_i \log \left(\left(\frac{1}{1 + e^{-x_i\beta}} \right)^{Y_i} \left(\frac{1}{1 + e^{x_i\beta}} \right)^{1-Y_i} \right) \quad (17)$$

$$= \sum_i \log \left(\frac{1}{1 + e^{-x_i\beta}} \right)^{Y_i} + \log \left(\frac{1}{1 + e^{x_i\beta}} \right)^{1-Y_i} \quad (18)$$

$$= \sum_i Y_i x_i \beta - \log(1 + e^{-x_i \beta}). \quad (19)$$

Because of the high-dimensional feature of our problem, we have to control the number of explanatory variables included in the model. Hence, an L_1 penalty a.k.a LASSO penalty is added to the log-likelihood as a penalty to including more explanatory variables in the model. Logistic regression with LASSO minimizes the negative logistic log-likelihood (4) with a Lasso penalty as below

$$\min - \sum_{t=1}^N [Y_i x_i \beta - \log(1 + e^{x_i \beta})] + \lambda |\beta|_1. \quad (20)$$

A well-known package called *glmnet* package provided by Hastie and Qian uses a quadratic approximation to the log-likelihood, and then coordinate descent on the resulting penalized weighted least-squares problem. And it is so far the most trust-worthy package in R for logistic regression with LASSO. For binary classification, we use the estimated β to construct a logistic probability model for y . Then, get our prediction from the model. If $\hat{\pi}(x) > 0.5$, the predicted class will be 1. And if $\hat{\pi}(x) < 0.5$, the predicted class will be 0.

We can use the following command for logistic regression.

```
library(JOUSBoost)
set.seed(111)
dat <- circle_data(n = 100)
x <- dat$X
y <- dat$y > 0
sum.data <- data.frame(x,y)

library(glmnet)
model <- cv.glmnet(y,x, family = "binomial")
y.fit <- predict(model, newx = x, s = "lambda.1se", type =
"response") > 0.5
train.error <- print(sum(y.fit != y) / n)
```

The output is the in-sample training error rate.

```
[1] 0.086
```

If we are interested in knowing the estimated coefficients of the model, we can check the components in the *model* object.

```
# The value of Lambda's as shown in the objective function
print(model$glmnet.fit$lambda)
# The estimated beta's using the corresponding lambda
# shown by the previous command
print(model$glmnet.fit$beta)
```

The output looks like this.

```
[1] 0.029100660 0.026515438 0.024159880 0.022013582
    0.020057956 0.018276063 0.016652468
[8] 0.015173109 0.013825171 0.012596981 0.011477900
    0.010458235 0.009529154 0.008682611
```

```
[15] 0.007911271 0.007208456 0.006568076 0.005984587
     0.005452932 0.004968509 0.004527120
[22] 0.004124943 0.003758495 0.003424601 0.003120368
     0.002843164 0.002590585
2 x 27 sparse Matrix of class "dgCMatrix"
[[ suppressing 27 column names s0, s1, s2 ... ]]

V1 . -0.0006865767 -0.001312223 -0.001882394 -0.00240204
     -0.0028939228 -0.0033458069 -0.003757714
V2 . .
  0.0003774458     0.0007948127  0.001175215

V1 -0.004133202 -0.004475500 -0.004787549 -0.005072024
     -0.005331360 -0.005567777 -0.005783296
V2  0.001521948  0.001838004  0.002126106  0.002388728
     0.002628126  0.002846351  0.003045275

V1 -0.005979761 -0.006158854 -0.006322106 -0.006470914
     -0.006606555 -0.006730189 -0.006842877
V2  0.003226601  0.003391885  0.003542542  0.003679864
     0.003805030  0.003919112  0.004023090

V1 -0.006945586 -0.007039197 -0.007124515 -0.007202272
     -0.007273138
V2  0.004117857  0.004204228  0.004282945  0.004354685
     0.004420064
```

Note that I reformat the output to fit the size of the chapter.

4.3 Semiparametric single-index model

[Chu et al. \(2018b\)](#) consider a standard single-index model,

$$y = m(x'\beta) + u, \quad (21)$$

where $\beta = (\beta_1, \dots, \beta_k)$ is a vector of coefficients. Under the sparsity condition, we assume that $\beta_j \neq 0$ for $j \leq r$ and $\beta_j = 0$ for $j > r$. We also assume that the random errors u are independent. However, we allow the presence of heteroskedasticity to encompass a large category of models for binary prediction, e.g. Logit and Probit models. The kernel estimator ([Ichimura, 1993](#)) we use is as shown below

$$\hat{m}(x'\beta; h) = \frac{\sum_{i=1}^n y_i K\left(\frac{X'_i \beta - x' \beta}{h}\right)}{\sum_{i=1}^n K\left(\frac{X'_i \beta - x' \beta}{h}\right)}, \quad (22)$$

where $K(\cdot)$ is a kernel function. The semiparametric kernel regression looks for the best β and h to minimize a weighted squared error loss. However,

exact identification is not available. If one blows up β and θ simultaneously by multiplying the same constant, the kernel estimator would yield identical estimates and losses. The standard identification approach is to set the first element of β to be 1 (Ichimura, 1993).

In terms of variable selection and prediction, we only need to focus on finding the best $\theta \equiv \frac{\beta}{h}$. Hence, we can simplify the estimator to

$$\hat{m}(x'\theta) = \frac{\sum_{i=1}^n y_i K(X'_i \theta - x' \theta)}{\sum_{i=1}^n K(X'_i \theta - x' \theta)}. \quad (23)$$

The basic idea of the SIM-Rodeo is to view the local bandwidth selection as a variable selection in sparse semiparametric single-index model. The SIM-Rodeo algorithm amplifies the inverse of the bandwidths for relevant variables while keeping the inverse of the bandwidths of irrelevant variables relatively small. The SIM-Rodeo algorithm is greedy as it solves for the locally optimal path choice at each iteration. It can also be shown to attain the consistency in mean square error when it is applied for sparse semiparametric single-index models. SIM-Rodeo is able to distinguish truly relevant explanatory variables from noisy irrelevant variables and gives a consistent estimator of the regression function. In addition, the algorithm is fast to finish the greedy steps.

Now we derive the Rodeo for Single-Index Models. First we introduce some notation. Let

$$W_x = \begin{pmatrix} K(X'_1 \theta - x' \theta) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & K(X'_n \theta - x' \theta) \end{pmatrix} \quad (24)$$

where $K(\cdot)$ is the Gaussian kernel. The standard Ichimura (1993) estimator takes the form

$$\hat{m}(x'\theta) = \frac{\sum_{i=1}^n y_i K(X'_i \theta - x' \theta)}{\sum_{i=1}^n K(X'_i \theta - x' \theta)} = (l' W_x l)^{-1} l' W_x y. \quad (25)$$

The derivative of the estimator Z_j with respect to θ_j is

$$Z_j \equiv \frac{\partial \hat{m}(x'\theta)}{\partial \theta_j} \quad (26)$$

$$\begin{aligned} &= (l' W_x l)^{-1} l' \frac{\partial W_x}{\partial \theta_j} y - (l' W_x l)^{-1} l' \frac{\partial W_x}{\partial \theta_j} l (l' W_x l)^{-1} l' W_x y \\ &= (l' W_x l)^{-1} l' \frac{\partial W_x}{\partial \theta_j} (y - l \hat{m}(x'\theta)). \end{aligned} \quad (27)$$

For the ease of computation, let

$$L_j = \begin{pmatrix} \frac{\partial \log K(X'_1 \theta - x' \theta)}{\partial \theta_j} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{\partial \log K(X'_n \theta - x' \theta)}{\partial \theta_j} \end{pmatrix}. \quad (28)$$

Note that

$$\frac{\partial W_x}{\partial \theta_j} = W_x L_j, \quad (29)$$

which appears in Eq. (27). With the Gaussian kernel, $K(t) = e^{-\frac{t^2}{2}}$, then L_j becomes

$$\begin{aligned} L_j &= \begin{pmatrix} -\frac{1}{2} \frac{\partial (X'_1 \theta - x' \theta)^2}{\partial \theta_j} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & -\frac{1}{2} \frac{\partial (X'_n \theta - x' \theta)^2}{\partial \theta_j} \end{pmatrix} \\ &= \begin{pmatrix} -(X'_1 \theta - x' \theta)(X_{1j} - x_j) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & -(X'_n \theta - x' \theta)(X_{nj} - x_j) \end{pmatrix}, \end{aligned}$$

where X_{1j} and X_{nj} are the j th elements of vectors X_1 and X_n . And x_j is the j th element of vector x . To simplify the notation, let $B_x = (l' W_x l)^{-1} l' W_x$. Then, the derivative Z_j becomes

$$\begin{aligned} Z_j &= (l' W_x l)^{-1} l' \frac{\partial W_x}{\partial \theta_j} (y - m(x' \theta)) \\ &= B_x L_j (I - l B_x) y \\ &\equiv G_j(x, \theta) y. \end{aligned} \quad (30)$$

Next, we give the conditional expectation and variance of Z_j .

$$Z_j = G_j(x, \theta) y = G_j(x, \theta) (m(x' \beta) + u), \quad (31)$$

$$E(Z_j | X) = E(G_j(x, \theta) (m(x' \beta) + u) | X) = G_j(x, \theta) m(x' \beta), \quad (32)$$

$$\text{Var}(Z_j | X) = \text{Var}(G_j(x, \theta) (m(x' \beta) + u) | X) = \boldsymbol{\sigma}' G_j(x, \theta)' G_j(x, \theta) \boldsymbol{\sigma}, \quad (33)$$

where $\boldsymbol{\sigma} = (\sigma(u_1), \dots, \sigma(u_n))'$ is the vector of standard deviations of u . In the algorithm, it is necessary to insert an estimate of σ . Since we allow the errors

to be heteroskedastic as in Logit and Probit models and estimate $\sigma(u_i)$ using the estimator $\hat{\sigma}(u_i) = m(x'_i\hat{\theta})(1 - m(x'_i\hat{\theta}))$.

SIM-Rodeo is described in [Algorithm 5](#), which is a modified algorithm of Rodeo ([Lafferty and Wasserman, 2008](#)).

We start by setting $\theta_j = \theta_0$ that is close to zero. Hence, $(X'_i\theta - x'\theta)$ are close to zero and $K(X'_i\theta - x'\theta)$ are close to $K(0)$. This means our estimator starts with the simple average of all observations, \bar{y} . If the derivative of θ_j is statistically different from zero. We amplify θ_j . If x_j is indeed a relevant explanatory variable, then the weights $K(X'_i\theta - x'\theta)$ change according to x_j . The estimator will give higher weights to observations close to $x'\theta$ and lower weights to observations away from $x'\theta$.

5 Monte Carlo

In this section, we demonstrate the above DAB, RAD, LB, and GB via small Monte Carlo simulation designs to illustrate R functions and library.

We construct the two DGPs to check the finite sample properties of the Boosting algorithms. DGP1 is a binary logistic model where y follows a Bernoulli distribution with probability

$$\pi(x) \equiv \frac{1}{1 + e^{-x\beta}}$$

ALGORITHM 5 SIM-Rodeo ([Chu et al., 2018b](#))

1. Select a constant $0 < \alpha < 1$ and the initial value

$$\theta_0 = c_0 \log \log n$$

where c_0 is sufficiently small. Compute Z_j with $\theta_j = \theta_0$ for all j .

2. Initialize the coefficients θ , and activate all covariates:

$$(a) \quad \theta_j = \begin{cases} \theta_0 & Z_j > 0 \\ -\theta_0 & \text{otherwise,} \end{cases} \quad j = 1, \dots, k.$$

$$(b) \quad \mathcal{A} = \{1, \dots, k\}.$$

3. While $\mathcal{A} \neq \emptyset$ is nonempty, do for each $j \in \mathcal{A}$:

(a) Compute Z_j and $s_j = \sqrt{\text{Var}(Z_j|X)}$ using (30) and (33) respectively.

(b) Compute the threshold $\lambda_j = s_j \sqrt{2 \log n}$.

(c) If $|Z_j| > \lambda_j$, then set $\theta_j \leftarrow \frac{\theta_j}{\alpha}$; Otherwise, remove j from \mathcal{A} (i.e., $\mathcal{A} \leftarrow \mathcal{A} - \{j\}$).

4. Obtain $\hat{\theta} = (\theta_1, \dots, \theta_k)$. Output the class probability prediction $\hat{\pi}(x) = \hat{m}(x'\hat{\theta})$ and the classifier $F(x) = 1_{(\hat{\pi}(x) > 0.5)}$.

to be 1 and $1 - \pi(x)$ to be -1 where

$$x_{n \times k} \sim N\left(0, \frac{\Sigma}{\beta' \Sigma \beta}\right), \quad \Sigma_{ij} = \rho^{|i-j|},$$

$$n = 100, k = \{2, 20\} \text{ and } \rho \in \{0\}.$$

We have two settings for the β . In the low-dimension case ($k = 2$), we let

$$\beta = (1, 1).$$

In the high-dimension case ($k = 20$), we let $\beta = (\beta_1, \dots, \beta_k)$ where

$$\beta_i = 0.9^i. \quad (34)$$

that decrease exponentially. Hence, most of the β 's are very close 0.

DGP2 is the circle model introduced in [Section 2](#). Here we have two settings for the circle model. In the low-dimension case, only the two relevant x 's are used to train the models as shown in the toy demo in previous sections. In the high-dimension (sparse) case, three irrelevant x 's are added in addition to the two relevant ones. [Tables 1](#) and [2](#) show the results of DGP2 (circle model) for the in-sample training error and the out-of-sample test error in the two cases (low-dimension and high-dimension) for seven different methods.

To construct the training and testing samples, we randomly generate x using the above distribution and calculate $\pi(x)$. To generate the random variable y based on x , we first generate a random variable ϵ that follows uniform distribution between $[0, 1]$. Next, we compare ϵ with $\pi(x)$. There is a probability of $\pi(x)$ that ϵ is smaller than $\pi(x)$ and a probability $1 - \pi(x)$ otherwise. Hence, we set

$$y = \begin{cases} 1 & \epsilon < \pi(x) \\ -1 & \epsilon > \pi(x). \end{cases}$$

TABLE 1 Error rate of low-dimension circle model

| | Train error | Test error |
|---------------------|-------------|------------|
| Discrete AdaBoost | 0.0820 | 0.2053 |
| Real AdaBoost | 0.0853 | 0.2038 |
| LogitBoost | 0.0602 | 0.2090 |
| Gentle AdaBoost | 0.0718 | 0.2062 |
| Deep Neural Network | 0.2601 | 0.3533 |
| Logistic Regression | 0.3586 | 0.3573 |
| SIM-RODEO | 0.2986 | 0.3421 |

TABLE 2 Error rate of high-dimension (sparse) circle model

| | Train error | Test error |
|---------------------|-------------|------------|
| Discrete AdaBoost | 0.0202 | 0.2203 |
| Real AdaBoost | 0.0295 | 0.2165 |
| LogitBoost | 0.0081 | 0.2232 |
| Gentle AdaBoost | 0.0133 | 0.2208 |
| Deep Neural Network | 0.2838 | 0.4017 |
| Logistic Regression | 0.3569 | 0.3572 |
| SIM-RODEO | 0.3542 | 0.3541 |

Given a set of observations $\{(x, y)\}$, we compare the average loss (classification error) achieved by using different methods. The formula for the average loss is as below.

$$\text{ErrorRate} = \frac{1}{n} \sum 1(y_i \neq \text{sign}(F_M(x_i))), \quad (35)$$

where n is the number of observations in the set.

To evaluate the algorithms, first we train our predictors with the training data of size $n = 100$. Then, we use a testing data set that contains 100 new observations of (x, y) to compute the average loss (35) achieved by the Boosting algorithms, Deep Neural Network, Logistic Regression and semiparametric Single-Index Model for out-of-sample evaluations. The boosting algorithms are component-wise versions of the four methods as shown before. The alternative methods we have, Deep Neural Network, Logistic Regression with LASSO penalty and semiparametric single-index model with SIM-RODEO considers all variables at the same time. The number of Monte Carlo repetition for each DGP is 1000.

The results for DGP1 (logistic model) are shown below (Tables 3 and 4).

From the simulation results, we can see that the four boosting methods work well in both the circle model and the logistic model. LogitBoost has the smallest training error among all four boosting algorithms as well as the largest testing error. On the other hand, Real AdaBoost has the largest training error as well as the smallest testing error. Similar rules apply to the other two boosting methods. Smaller training errors imply larger testing errors. This is an evidence of overfitting which is related to the hyper-parameters in the boosting algorithms. If the number of boosting iterations is small, then we will have a larger training error but less risk of overfitting. On the other hand, if we have more boosting iterations, then the boosting methods will fit the

TABLE 3 Error rate of low-dimension logistic model

| | Train error | Test error |
|---------------------|-------------|------------|
| Discrete AdaBoost | 0.1431 | 0.3129 |
| Real AdaBoost | 0.1519 | 0.3120 |
| LogitBoost | 0.1302 | 0.3160 |
| Gentle AdaBoost | 0.1339 | 0.3154 |
| Deep Neural Network | 0.2304 | 0.3090 |
| Logistic Regression | 0.2773 | 0.3083 |
| SIM-RODEO | 0.3069 | 0.3415 |

TABLE 4 Error rate of high-dimension (sparse) logistic model

| | Train error | Test error |
|---------------------|-------------|------------|
| Discrete AdaBoost | 0.0007 | 0.3217 |
| Real AdaBoost | 0.0015 | 0.3215 |
| LogitBoost | 0.00007 | 0.3214 |
| Gentle AdaBoost | 0.0001 | 0.3204 |
| Deep Neural Network | 0.0523 | 0.3172 |
| Logistic Regression | 0.2328 | 0.3432 |
| SIM-RODEO | 0.3580 | 0.3971 |

training data better but raise higher risk on overfitting. The number of iterations in the boosting algorithms is fixed by the users. However, cross-validation could be used to determine the optimal number of iterations.

As for the alternative methods, Deep Neural Network works better in the logistic model than the circle model. This is a result of the setup of the Deep Neural Network. We use the logistic function as the activation function and output function, and the entropy as the loss function. The setup will give better results when logistic model is the true model. For the circle model, Deep Neural Network gives a comparable result to the Logistic Regression in the low-dimension case. However, the result is much worse for the high-dimension case. Again, this could be a result of our setup of the Deep Neural Network. We acknowledge that the Deep Neural Network is high flexible with lots of hyper-parameters Different setup of the model may lead to

dramatically distinct results. Our setting by no means is the optimal one and Deep Neural Network could perform better with a different setup.

For Logistic Regression, it works best in the low-dimension logistic model as all parametric assumptions are satisfied. However, in the high-dimension case, Logistic Regression will have a larger bias due to the need to shrink the coefficients of irrelevant variables to zero. To fix this bias, one may try the De-biased Machine Learning method (Chernozhukov et al., 2018).

6 Applications

In this section we illustrate the R functions in economics applications.

In the application, we use the FRED monthly data <https://research.stlouisfed.org/econ/mccracken/fred-databases/> to predict the moving direction of real personal income in the United States as in Ng (2014). After removing the observations with missing values, our obtain 341 effective observations with a sample period starting from September, 1989 to January 2018. We use 125 variables which are all variables in the data except for the Consumer Sentiment Index that is only available quarterly and New Orders for Consumer Goods which has too many missing data. We generate the direction of the real personal income as our dependent variable and take the lag of the dependent variable as one explanatory variable. Hence, we have in total $k = 126$ explanatory variables and $(341 - 1 = 340)$ observations. We use rolling training sample with window width $W = 100$ and predict the one month ahead moving direction. We have $(n = 340 - W = 240)$ subsamples and predictions. The results are shown in Table 5.

The results are very similar to the simulation results for logistic models. The boosting methods have very small in-sample training errors. However, the out-of-sample testing error is much larger than the alternatives. This may indicate that the boosting algorithms are overfitting the model.

TABLE 5 Error rate of application

| | Train error | Test error |
|---------------------|-------------|------------|
| Discrete AdaBoost | 0.0028 | 0.3125 |
| Real AdaBoost | 0.0407 | 0.3291 |
| LogitBoost | 0.0003 | 0.3041 |
| Gentle AdaBoost | 0.0020 | 0.3083 |
| Deep Neural Network | 0.2389 | 0.2666 |
| Logistic Regression | 0.2479 | 0.2708 |
| SIM-RODEO | 0.2257 | 0.2958 |

7 Conclusions

This chapter shows recent developed methods for high-dimensional binary classification and probability prediction. We start by introducing four component-wise boosting methods, namely component-wise Discrete AdaBoost, component-wise Real AdaBoost, component-wise LogitBoost, and component-wise Gentle AdaBoost. Discrete AdaBoost, Real AdaBoost, and Gentle AdaBoost minimizes the exponential loss via Newton-like procedures. LogitBoost minimizes the Bernoulli log-likelihood via adaptive Newton method. These methods are extremely popular since they are both computationally efficient and easy to implement. Moreover, the component-wise Boosting algorithms deal with high-dimensional issue by considering the explanatory one at a time. In each iteration, only the most effective explanatory variable is chosen to train a weak learner. Hence, these methods allows $k \gg n$. However, hyper-parameters such as the number of boosting iteration normally need to be determined by the user prior to the estimation procedure. Cross-validation may also be used to choose the number of iterations.

Next, we give an introduction to alternative methods such as Deep Neural Network, Logistic Regression, and SIM-RODEO. Deep Neural Network is a kind of nonlinear statistical learning model features a network structure that is similar to the relationship between the neurons of human brain. Deep Neural Network may be explained partly as a kind of basis transformation which leads to extreme flexibility of the model. Deep Neural Network and its variants are the most popular prediction method at this time and are widely used in fields such as image and voice recognition.

Logistic Regression is a traditional method used intensively in economics for binary classification and probability prediction. Logistic Regression assumes that the probability that the output label is 1 conditional on x follows a logistic function of x . Under such assumption, the parameters of the model often have practical economic meaning unlike machine learning methods that are often hard to interpret. However, logistic regression relies heavily on its parametric assumptions and is the least flexible model introduced in this chapter. In addition, to deal with high-dimensional problem, we have to use the LASSO to control the number of explanatory variables chosen in the model.

SIM-RODEO relaxes the parametric assumption of Logistic Regression. As a result, SIM-RODEO is more flexible but, to some extent, still interpretable as Logistic Regression. However, the flexibility of SIM-RODEO may lead to a slower convergence rate and less time efficiency.

This chapter conducted extensive comparison of the above mentioned methods through Monte Carlo experiments. We compare the methods using both traditional binary classification model (logistic model) and irregular model (circle model). The boosting methods work well in both the traditional models and irregular models. Logistic Regression works better in the low-dimension logistic model when the parametric assumptions of Logistic Regression are satisfied. However, in the high-dimensional case, the LASSO introduces high bias in Logistic Regression and lead to lower classification accuracy. In the irregular

models, Logistic Regression performs poor compared to the boosting algorithms. The Deep Neural Network performed best in the traditional methods as a result of our configuration of the Neural Network. We acknowledge that our configuration of Deep Neural Network is by no means the best and the results here may improve with different activation function, output function and/or number of hidden layers and neurons. SIM-RODEO is an extension to parametric methods such as Logistic Regression. It performs reasonably well in the models.

We also use these methods for predicting the changing direction of the real personal income in the United States. The application show similar results as in the simulation of logistic models.

This chapter gives a thorough introduction of newly developed methods for binary classification and probability prediction. Advantages and disadvantages of each method are discussed and compared. We conclude that no single method has an absolute advantage in all aspects over the other methods. We believe binary classification and probability prediction will remain important for business and economics and look forward to future works on this problem.

Acknowledgments

The authors would like to thank the editors, a referee, and Shujie Ma for helpful comments.

References

- Allaire, J., Chollet, F., 2018. keras: R Interface to ‘Keras’. R package version 2.1.6, <https://CRAN.R-project.org/package=keras>.
- Bliss, C.I., 1934. The method of probits. *Science* 79 (2037), 38–39. ISSN 00368075. <https://doi.org/10.1126/science.79.2037.38>. <http://www.ncbi.nlm.nih.gov/pubmed/17813446>.
- Bühlmann, P., 2006. Boosting for high-dimensional linear models. *Ann. Stat.* 34 (2), 559–583. ISSN 0900-5364. <https://doi.org/10.1214/009053606000000092>.
- Bühlmann, P., Yu, B., 2003. Boosting with the L_2 loss: regression and classification. *J. Am. Stat. Assoc.* 98 (462), 324–339. ISSN 01621459. <https://doi.org/10.1198/016214503000125>. <http://www.tandfonline.com/doi/abs/10.1198/016214503000125>.
- Chatterjee, S., 2016. fastAdaboost: A Fast Implementation of Adaboost. R package version 1.0.0. <https://CRAN.R-project.org/package=fastAdaboost>.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. *Econ. J.* 21 (1), C1–C68. ISSN 1368423X. <https://doi.org/10.1111/ectj.12097>. <http://arxiv.org/abs/1608.00060>.
- Chu, J., Lee, T.-H., Ullah, A., 2018a. Asymmetric AdaBoost for High-Dimensional Maximum Score Regression. University of California, Riverside.
- Chu, J., Lee, T.-H., Ullah, A., 2018b. Variable selection in sparse semiparametric single index model. In: Jeliazkov, I., Tobias, J.L. (Eds.), *Advances in Econometrics*, vol. 40. Emerald Group Publishers (forthcoming).
- Cox, D.R., 1958. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B Methodol.* 20 (2), 215–242. <http://www.jstor.org/stable/2983890>.
- Culp, M., Johnson, K., Michailidis, G., 2016. ada: The R Package ada for Stochastic Boosting. <https://CRAN.R-project.org/package=ada>. R package version 2.0-5.
- Elliott, G., Lieli, R.P., 2013. Predicting binary outcomes. *J. Econ.* 174 (1), 15–26. ISSN 03044076. <https://doi.org/10.1016/j.jeconom.2013.01.003>.

- Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55, 119–139.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. <https://doi.org/10.2307/2699986>.
- Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. *Ann. Stat.* 28 (2), 337–407. ISSN 00905364. <https://doi.org/10.1214/aos/1016218223>.
- Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Soft.* 33 (1), 1–22. ISSN 1548-7660. <https://doi.org/10.18637/jss.v033.i01>. <http://www.jstatsoft.org/v33/i01/>.
- Fritsch, S., Guenther, F., 2016. neuralnet: Training of Neural Networks. R package version 1.33, <https://CRAN.R-project.org/package=neuralnet>.
- Ichimura, H., 1993. Semiparametric least squares and weighted SLS estimation of single index models. *J. Econ.* 58 (1–2), 71–120. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.612.5853&rep=rep1&type=pdf>.
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An Introduction to Statistical Learning with Applications in R*. Springer.
- Lafferty, J., Wasserman, L., 2008. Rodeo: sparse, greedy nonparametric regression. *Ann. Stat.* 36 (1), 28–63. ISSN 00905364. <https://doi.org/10.1214/009053607000000811>.
- Lahiri, K., Yang, L., 2012. Forecasting binary outcomes. In: Elliott, G., Timmermann, A. (Eds.), *Handbook of Economic Forecasting*, vol. 2. SSRN, pp. 1025–1106.
- Manski, C.F., 1975. Maximum score estimation of the stochastic utility model of choice. *J. Econ.* 3 (3), 205–228. ISSN 03044076. [https://doi.org/10.1016/0304-4076\(75\)90032-9](https://doi.org/10.1016/0304-4076(75)90032-9). <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.587.6474&rep=rep1&type=pdf>.
- Manski, C.F., 1985. Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator. *J. Econ.* 27 (3), 313–333. ISSN 03044076. [https://doi.org/10.1016/0304-4076\(85\)90009-0](https://doi.org/10.1016/0304-4076(85)90009-0). <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.504.7329&rep=rep1&type=pdf>.
- Mease, D., Wyner, A., Buja, A., 2007. Cost-weighted boosting with jittering and over/under-sampling: jous-boost. *J. Mach. Learn. Res.* 8, 409–439.
- Ng, S., 2014. Viewpoint: boosting recessions. *Can. J. Econ.* 47 (1), 1–34. <https://doi.org/10.1111/caje.12070>.
- Olson, M., 2017. JOUSBoost: Implements Under/Oversampling for Probability Estimation. R package version 2.1.0, <https://CRAN.R-project.org/package=JOUSBoost>.
- Ridgeway, G., 2017. gbm: Generalized Boosted Regression Models. R package version 2.1.3, <https://CRAN.R-project.org/package=gbm>.
- Su, L., Zhang, Y., 2014. Variable selection in nonparametric and semiparametric regression models. In: Racine, J.S., Liangjun, S., Ullah, A. (Eds.), *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Oxford University Press. <http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199857944.001.0001/oxfordhb-9780199857944-e-009>.
- Tibshirani, R., 1996. Regression selection and shrinkage via the lasso. *J. R. Stat. Soc. B* 58 (1), 267–288. ISSN 00359246. <https://doi.org/10.2307/2346178>. 11/73273, <http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574>.
- Tuszynski, J., 2018. caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.17.1.1, <https://CRAN.R-project.org/package=caTools>.
- Walker, S.H., Duncan, D.B., 1967. Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54 (1), 167–179. ISSN 00063444. <https://doi.org/10.1093/biomet/54.1-2.167>.
- Zou, H., 2006. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* 101 (476), 1418–1429. <https://doi.org/10.1198/016214506000000735>.

Chapter 4

Mixed data sampling (MIDAS) regression models

Eric Ghysels^{a,b,*}, Virmantas Kvedaras^c and Vaidotas Zemlys-Balevičius^d

^a*Department of Economics, UNC Chapel Hill, Chapel Hill, NC, United States*

^b*Department of Finance, Kenan-Flagler Business School and CEPR Fellow, Chapel Hill, NC, United States*

^c*Joint Research Centre, European Commission, Brussels, Belgium*

^d*Institute of Applied Mathematics, Vilnius University, Vilnius, Lithuania*

*Corresponding author: e-mail: eghsels@unc.edu

Abstract

Mixed data sampling (MIDAS) regressions are now commonly used to deal with time series data sampled at different frequencies. This chapter focuses on single-equation MIDAS regression models involving stationary processes with the dependent variable observed at a lower frequency than the explanatory ones. We discuss in detail nonlinear and semiparametric MIDAS regression models, topics not covered in prior work. Moreover, fitting the theme of the handbook, we also elaborate on the R package `midasr` associated with the regression models using simulated and empirical examples. In the theory part, a stylized model is introduced in order to discuss specific issues relevant to the construction of MIDAS models, such as the use or nonuse of functional constraints on parameters, the types of constraints and their choice, and the selection of the lag order. We introduce various new MIDAS regression models, including quasi-linear MIDAS, models with nonparametric smoothing of weights, logistic smooth transition and min–mean–max effects MIDAS, and semiparametric specifications.

Keywords: MIDAS regressions, Mixed frequency data, Semiparametric regression, Smooth transition models, Index models

1 Introduction

Raw economic data features a wide range of sampling frequencies. Financial data are typically available without revisions daily or even intradaily, macrodata are with a few exceptions usually observed monthly or quarterly and are often

^{*}The opinions expressed are those of the authors only and should not be considered as representative of the European Commission's official position.

revised. This frequency imbalance has spurred in recent years a literature dealing explicitly with models that tackle the issue. One of the popular approaches is mixed data sampling (MIDAS) regressions and related econometric methods.

In a baseline MIDAS regression deals with a single low-frequency variable projected onto a high-frequency variable—possibly augmented with the lagged-dependent variable. When the difference in sampling frequencies is small, we may treat this as a regular regression problem, now dubbed the unconstrained MIDAS model (U-MIDAS, see [Foroni et al., 2015](#)) namely when dealing with, for instance, yearly projected onto quarterly—or quarterly projected onto monthly.

When the difference in observation frequency increases, the number of parameters associated with each high-frequency lag grows and the use of U-MIDAS becomes either unappealing or even infeasible. Therefore, one faces a challenging problem of a regression model with a potentially large number of regressors—potentially even larger than the sample size. Since this is similar to machine learning, one might think that solutions such as shrinkage estimators based, e.g., on the least absolute shrinkage and selection operator (*LASSO*) or *Ridge* regression, or using other dimensionality reduction approaches, would work.

If only a few high-frequency observations were relevant in the regression, other high-frequency lags being uninformative even jointly, then the *LASSO* estimator would be suitable, as it effectively would shrink to zero all but a few high-frequency regressor lags. Mixed frequency regressions do not correspond to this sparsity setting as typically numerous individually small and similar contributions would be relevant jointly. While the L^2 regularization of *Ridge* regression might be more appealing, it is not well suited either when there is a pattern present of smooth decay of coefficients, which has been shown to be the usual situation in many time series regression applications. In such a situation some other dimension-reduction procedures would be preferable.

For cases involving a low-frequency-dependent variable, [Ghysels et al. \(2002\)](#) introduced, while [Ghysels et al. \(2006, 2007\)](#) further developed the MIDAS regression approach that has several distinctive features. First, it does not rely on preaggregated values, but instead uses the regression function to extract the best fitting alignment of low- and high-frequency data. Second, the frequency alignment is achieved by using a flexible and parsimonious parametric functional constraint of the impact of high-frequency variables on the low-frequency variables. The functional form is assumed to be known, whereas the parameter values of the function are unknown and therefore estimated. Third, there are many approaches to choosing the functional form and reducing the parameter dimension, prompting the need to select the proper functional form—in the context of MIDAS regressions (see, e.g., [Kvedaras and Račkauskas, 2010](#); [Kvedaras and Zemlys, 2012](#)). As both the regression function and the functional form of the true constraint might be unknown, the MIDAS regressions are often bearing the projection interpretation. Nevertheless, the regression interpretation is fully admissible at least in the [White \(1981\)](#) sense, i.e., as the best approximation of the underlying true regression by some constrained function.

There are a number of existing surveys covering a broad overview on various aspects of MIDAS modeling. [Armesto et al. \(2010\)](#) provide a layman's introduction to MIDAS regressions, whereas [Andreou et al. \(2011\)](#), [Foroni and Marcellino \(2013\)](#), and [Ghysels et al. \(2016\)](#) cover MIDAS in a more general setting. [Ghysels \(2013\)](#) presents the MATLAB MIDAS Toolbox, while [Ghysels and Valkanov \(2012\)](#) discuss volatility models and MIDAS.

This chapter focuses on single-equation MIDAS regression models involving stationary processes with the dependent variable observed at a lower frequency than the explanatory ones. We do not discuss systems of equations such as, e.g., mixed frequency VAR models (see [Ghysels, 2016](#)). In addition, we confine our attention to classical estimators (see, e.g., [Rodriguez and Puggioni, 2010](#), for a discussion of the Bayesian estimation of MIDAS models). However, we do discuss in detail nonlinear and semiparametric MIDAS regression models, topics not covered in prior survey. Moreover, fitting the theme of the handbook, we also elaborate the R code associated with the regression models.

It is worth noting that state space (SS) models have also been used to deal with mixed frequency data—usually estimated using the Kalman filter (see [Bai et al., 2013](#) for a comparison). A SS formulation allows one to impose further structure on the relationships between mixed frequency data in comparison to MIDAS regressions. In fact, MIDAS regressions can be thought of as some reduced-form prediction formula representation which emerges from the SS model and Kalman filter. For nonlinear and nonparametric models, there is no straightforward SS counterpart and, especially, their estimation becomes much harder or even impossible (see, e.g., [Teräsvirta et al., 2010](#), chapter 9), whereas the MIDAS regression is easy to extend to various forms of nonlinearity with simple applications of rather standard estimators. In this chapter, we will concentrate on a few nonlinear formulations that can be estimated using standard or slightly extended nonlinear least squares (NLS) estimators. This will also include some semiparametric variants of MIDAS regressions. We do not consider fully nonparametric estimation (see [Breitung and Roling, 2015](#)).

The remainder of this chapter has the following structure: (1) the theoretical models are presented and then (2) using the R package `midasr` simulated and empirical examples are used to illustrate their performance. In the theory part, a stylized model is introduced in order to discuss specific issues relevant to the construction of MIDAS models, such as the use or nonuse of functional constraints on parameters, the types of constraints and their choice, the selection of the lag order, etc.

2 A stylized MIDAS regression model

Let $t \in \mathbb{N}$ index the low-frequency observations of a dependent variable $y_t \in \mathbb{R}$, and $\tau \in \mathbb{N}$ for indexing the high-frequency observations. For simplicity of presentation, we will use a single high-frequency explanatory variable $x_\tau \in \mathbb{R}$, which is easily extendable to the case of many high-frequency variables having various frequencies and various constraints. It should be pointed out that

there is no requirement in MIDAS (regression) models to have a fixed number of high-frequency observations per a low-frequency period. Hence, let m_t denote the number of high-frequency observations pertaining to the t th low-frequency observation. Consequently, the total number of high-frequency periods available up till (and including) the t th low-frequency observation is given by $s(t) = \sum_{j=1}^t m_j$, which by construction coincides with the index τ .

Let us focus on a case involving a single low- and high-frequency variable and consider the projection of the high-frequency series onto the low one. The stylized MIDAS regression model we consider is:

$$\begin{aligned} y_t &= g\left(\sum_{i=0}^k w_i x_{s(t)-i}; \boldsymbol{\beta}\right) + \varepsilon_t, \text{s.t.} \\ \forall i, w_i &= h(\boldsymbol{\gamma}, i) \quad \text{and} \quad \sum_{i=0}^k w_i = 1. \end{aligned} \tag{1}$$

The function $g: \mathbb{R} \rightarrow \mathbb{R}$ can be either parametric, in which case $\boldsymbol{\beta} \in \mathbb{R}^b$ is its low-dimensional parameter vector (b here is some small positive integer). Moreover, when $g(z; \boldsymbol{\beta}) = g(z) = z \quad \forall z \in \mathbb{R}$ we obtain the model dubbed by Andreou et al. (2011, 2013) as the DL-MIDAS regression. For nonparametric g , $\boldsymbol{\beta} \equiv 1$ will be imposed. The key feature of MIDAS regressions is the alignment of low- and high-frequency variables through a functional constraint $h: \mathbb{R}^d \times \mathbb{N} \rightarrow \mathbb{R}$, $d \in \mathbb{N}$, on the (impact of) high-frequency variable at lag i using weights $w_i = h(\boldsymbol{\gamma}, i)$, involving the low-dimensional parameter vector $\boldsymbol{\gamma} \in \mathbb{R}^d$, and the lag index $i \in \{0, 1, \dots, k\}$.

The zero mean error term ε_t is independent of the high-frequency explanatory variable and, for simplicity of presentation, also identically and independently distributed (i.i.d.), which to some degree can be relaxed using the usual approaches through adjusting the corresponding covariance matrices of estimators and/or properly transforming the original observations to obtain the spherical error term. The normalization condition $\sum_{i=0}^k w_i = 1$ is often required in (1) for the identification of $\boldsymbol{\beta}$ and/or g .

A couple of notes are worth pointing out regarding $\sum_{i=0}^k w_i x_{s(t)-i}$. First, it leads to the usual interpretation of aggregates when $\forall t, m_t > k$, because—assuming one knew the function h and its parameter vector $\boldsymbol{\gamma}$ —it would be possible to transform high-frequency data into low-frequency aggregates for each low-frequency period separately. In the general case, typically $m_t \leq k$, leading to a more general frequency alignment compared to standard aggregation. Second, since the parameter vector $\boldsymbol{\gamma}$ is estimated from the data, the constraint is flexible up to the shapes admissible by function h . Third, when the number of high-frequency periods per a low-frequency period is constant, i.e., $\forall t, m_t = m$ and $s(t) = m \cdot t$, other notations/representations are often employed in the MIDAS literature, namely:

$$\sum_{i=0}^k w_i x_{s(t)-i} \Big|_{\forall t, m_t = m} = \tilde{x}_{t(\boldsymbol{\gamma}, k)}^{(m)} = H(L^{1/m}; \boldsymbol{\gamma}) x_{mt} = \sum_{i=0}^k h(\boldsymbol{\gamma}, i) L^{i/m} x_{mt}, \tag{2}$$

where polynomial $H(L^{1/m}; \gamma) = \sum_{i=0}^k \gamma_i L^{i/m}$. It is clear that, for instance, monthly to daily and higher-frequency relations often violate the condition $\forall t, m_t = m$. Nevertheless, from a practical point of view, this is not very constraining, because one can take $m := \max_t m_t$. From the mathematical point of view, this would require introduction of additional selection matrices or indicator functions to deal with missing/nonmissing observations that would further obscure the notation and presentation, therefore we will mostly stick to the representation with fixed m .

2.1 A few examples of the constraint function h

Since w_i are restricted to add to one, it is convenient to represent the functional constraint h in the following form

$$w_i = h(\gamma, i) = \frac{\psi(\gamma, i)}{\sum_{j=0}^k \psi(\gamma, j)}. \quad (3)$$

Here the choice of the underlying function $\psi: \mathbb{R}^d \times \mathbb{N} \rightarrow \mathbb{R}$ determines the shape of h . In principle, any parametric function can be applied provided it achieves sufficient flexibility under a tight parameterization (low dimension d for vector γ) with $\sum_{j=0}^k \psi(\gamma, j) \neq 0$. For some results to be presented later, it will be assumed that h is twice differentiable.

The most widely used functions ψ in applied work are:

- *exponential Almon polynomials* with $\psi(\gamma, i) = \exp\left(\sum_{j=1}^d \gamma_j i^j\right)$, $d \in \{2, 3\}$;
- the *beta polynomial* with $\psi(\gamma, i) = x_i^{\gamma_1 - 1} (1 - x_i)^{\gamma_2 - 1}$,

where $x_i = \xi + (1 - \xi) \frac{i-1}{k-1}$ and marginally small quantity $\xi > 0$. In the beta polynomial $d=2$, although [Ghysels and Qian \(2019\)](#) suggest further imposing $\gamma_1 \equiv 1$ that leads to a still quite flexible constraint with only a single parameter γ_2 to be estimated, i.e., $d=1$ in such a case;

- and a *hyperbolic scheme polynomial* with $\psi(\gamma, i) = \frac{\Gamma(i+\gamma)}{\Gamma(i+1)\Gamma(\gamma)}$,

where $\Gamma(a) = \int_0^\infty e^{-x} x^{a-1} dx$. In this case $d=1$, as there is a single parameter γ to be estimated.

It is clear that these choices also ensure the nonnegativity of w_i , which might or might not be desirable depending on the modeled phenomenon. Many other alternatives can be designed with their taxonomy given in [Ghysels et al. \(2016\)](#) and additional examples further discussed in, e.g., [Ghysels \(2013\)](#).

2.2 Selection of h , d , and k

Leaving aside the usual considerations about the properties of the error term, there are two main questions about the specification of MIDAS regression models. First, a suitable functional constraint h needs to be selected (including

its complexity in terms of the number of parameters in γ , i.e., d), since their choice will constrain the achievable precision of the model. Second, the appropriate maximum lag order k needs to be chosen, with too-low order potentially being of greater concern, as the number of parameters in the constrained model does not change with larger k , and therefore having little influence whenever the constraint function h is adequate. Other than in the constrained models, where the selection of pair (h, d) is of greater importance, in the unconstrained U-MIDAS regressions there is no constraint on parameters and therefore the selection of k corresponds directly to the selection of number of parameters in this regression.

Provided there are a sufficiently large number of observations, the usual split of data into the estimation, selection, and evaluation subsamples could be used for the choice of h , d , and k from a list of alternative candidates. In many empirical applications, there is no abundance of data and therefore the usage of information criteria to select the best model in terms of the parameter restriction and the lag orders is chosen, which is notably encouraged by [Diebold and Mariano \(2002\)](#) instead of out-of-sample precision evaluation procedures in order to omit the loss of data due to the sample splitting. However, such an approach is feasible only for parametric models. Furthermore, one should be careful to use the proper likelihood in the information criteria depending on the model and the properties of the error term. For instance, in the standard functions described in [Ghysels et al. \(2016\)](#), the linear model with i.i.d. errors underlies the calculated information criteria—see, for example, [Foroni et al. \(2018\)](#) who show that certain aggregation schemes may introduce MA errors in MIDAS regressions. In addition, the information criteria-based approach is not feasible for semiparametric models to be considered later on. Finally, the cross-validation is a general approach that will be applicable in all the cases that will be under consideration.

It should be also pointed out that even the best choice of a pair (h, d) out of a constrained set of some potential candidates might not be empirically adequate. Hence, testing of adequacy of the functional constraint as suggested in [Kvedaras and Zemlys \(2012\)](#) seems to be of importance. The model adequacy becomes crucially important whenever the model is intended to be used for economic interpretation based on some statistical inference. Whenever the modeling aim is prediction, this might be of less concern as inadequacy would result in some under-performance of forecasts.

2.3 Statistical inference

In this section, we also characterize briefly the basic statistical inference principles to be used in the empirical applications relying on the theory to be presented in the following sections. Using the notation of the stylized MIDAS regression as presented in Eq. (1), let $\theta = (\beta', \gamma')'$. Suppose that its estimator $\hat{\theta}$ satisfies

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}) \quad (4)$$

with some covariance matrix \mathbf{V} , where T denotes the number of observations. This implies that

$$P\left\{\frac{z_\alpha}{2} < \frac{\sqrt{T}\mathbf{c}'(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})}{\sqrt{\mathbf{c}'\mathbf{V}\mathbf{c}}} < z_{1-\frac{\alpha}{2}}\right\} \rightarrow 1 - \alpha, \quad \alpha \in [0, 1]. \quad (5)$$

Here z_p denotes the p th percentile/quantile of the *normal* distribution, whereas \mathbf{c} is some vector of constants with the particular interest being in the selection vector \mathbf{c}_i taking value one for an i th parameter and zero otherwise.

In empirical applications, the confidence bands and significance of coefficients will be evaluated based on (5) using a few critical values ($\alpha \in \{0.01, 0.05, 0.1\}$ is usually under consideration) replacing the unknown covariance matrix \mathbf{V} with its consistent estimator $\hat{\mathbf{V}}$. Letting the standard error of an i th coefficient $\hat{\theta}_i = \mathbf{c}_i' \hat{\boldsymbol{\theta}}$ be denoted by s.e. $(\hat{\theta}_i) = \mathbf{c}_i' \hat{\mathbf{V}} \mathbf{c}_i / \sqrt{T}$, the coefficient $\hat{\theta}_i$ will be said to be statistically significant (from zero) at the α significance level, provided $|\hat{\theta}_i| / \text{s.e. } (\hat{\theta}_i) > z_{1-\frac{\alpha}{2}}$, whereas the (approximate) $1 - \alpha$ confidence bounds will be defined by $\hat{\theta}_i \pm z_{1-\frac{\alpha}{2}} \cdot \text{s.e. } (\hat{\theta}_i)$.

It should be further noted that from (4) similar results can be usually derived also for some differentiable function f of the parameter vector $\boldsymbol{\theta}$, because using the *delta method* it follows from (4) that $\sqrt{T}(f(\hat{\boldsymbol{\theta}}) - f(\boldsymbol{\theta})) \xrightarrow{d} N(\mathbf{0}, \partial f / \partial \boldsymbol{\theta}' \mathbf{V} (\partial f / \partial \boldsymbol{\theta}')')$ with the same implications for the construction of (approximate) confidence bounds relying on the consistent estimators $f(\hat{\boldsymbol{\theta}})$, $\hat{\mathbf{V}}$, and $\partial f / \partial \boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}}}$. The relevant examples for the MIDAS model are either the restriction function h in Eq. (1) or the regression function g itself.

3 Linear and quasi-linear MIDAS models (affine g)

In this section, we consider models with an affine g in Eq. (1), i.e., it is assumed that the regression function in terms of $\{x_{s(t)-i}\}_{i=0}^k$ is linear. Also the $\varepsilon_t \sim \text{i.i.d. } (0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon^2 \in \mathbb{R}_+$ will be retained. This leads to the following regression model

$$y_t = \beta_0 + \beta_1 \sum_{i=0}^k w_i x_{s(t)-i} + \varepsilon_t, \text{ s.t. } \forall i, w_i = h(\gamma, i) \text{ and } \sum_{i=0}^k w_i = 1 \quad (6)$$

which is the basic MIDAS regression model typically encountered in the literature. It should be pointed out that it is linear in terms of variables, but

not in terms of parameters, therefore requiring some nonlinear estimators whenever the restrictions are taken into account. It can be connected to a fully linear model ignoring the constraints, which is discussed next.

3.1 Unconstrained MIDAS

Let us define $b_i = \beta_1 w_i$, $i \in \{0, \dots, k\}$, considering each such coefficient as a separate parameter, i.e., ignoring that $\{w_i\}_{i=0}^k$ are generated by the function h . Then model (6) leads to a linear unconstrained MIDAS model (U-MIDAS) introduced by Foroni et al. (2015):

$$y_t = \beta_0 + \sum_{i=0}^k b_i x_{s(t)-i} + \epsilon_t, \quad (7)$$

which has $k+2$ unconstrained parameters $\boldsymbol{\beta}_u = (\beta_0, b_0, b_1, \dots, b_k)'$ to be estimated in the regression function. It is attractive, because, under the linearity of the regression function, the parameter vector can be consistently estimated by the ordinary least squares (OLS)

$$\hat{\boldsymbol{\beta}}_u = \underset{\beta_0, b_0, b_1, \dots, b_k}{\operatorname{argmin}} \sum_t \left(y_t - \beta_0 - \sum_{i=0}^k b_i x_{s(t)-i} \right)^2 = \left(\sum_{t=1}^T \mathbf{x}_{s(t)} \mathbf{x}'_{s(t)} \right)^{-1} \sum_{t=1}^T \mathbf{x}_{s(t)} y_t, \quad (8)$$

where $\mathbf{x}_{s(t)} := (1, x_{s(t)}, x_{s(t)-1}, \dots, x_{s(t)-k})'$ (see, e.g., Ghysels and Marcellino, 2018, chapter 1). In addition, under the usual regularity conditions

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_u - \boldsymbol{\beta}_u) \xrightarrow{d} N(\mathbf{0}_{k+2}, \mathbf{V}_u)$$

as $T \rightarrow \infty$, where $\mathbf{V}_u = \sigma_e^2 \operatorname{plim} \mathbf{S}_x^{-1}$, $\sigma_e^2 = E(\epsilon_t^2)$, and $\mathbf{S}_x = \left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_{s(t)} \mathbf{x}'_{s(t)} \right)$.

It is of interest to point out that, from the components of $\hat{\boldsymbol{\beta}}_u$, the point estimates of β_1 and w_i , $i \in \{0, \dots, k\}$ of model (6) can be obtained from

$$\hat{\beta}_1 = \sum_{i=0}^k \hat{b}_i, \quad \hat{w}_i = \frac{\hat{b}_i}{\sum_{j=0}^k \hat{b}_j}, \quad i \in \{0, \dots, k\},$$

with their variances $\operatorname{Var}(\hat{\beta}_1) = \sum_{i=0}^k \sum_{j=0}^k \operatorname{Cov}(\hat{b}_i, \hat{b}_j)$ and $\operatorname{Var}(w_i) = \operatorname{Var}(\hat{b}_i / \hat{\beta}_1)$, $i \in \{0, \dots, k\}$. The former one, being linear in terms of covariances, can be easily calculated, as the covariances $\operatorname{Cov}(\hat{b}_i, \hat{b}_j)$ are the elements of the triangular submatrix obtained from the estimated \mathbf{V}_u by dropping its first line and column pertaining to the intercept β_0 . The variances of $\{\hat{w}_i\}_{i=0}^k$ are more complicated to evaluate.

Apart from the linearity of the regression function, the U-MIDAS does not impose any constraint on its parameters, therefore there cannot emerge any bias due to a potentially incorrect restriction placed on coefficients $\{b_i\}_{i=0}^k$. Hence the U-MIDAS regression is a reasonable choice for modeling and forecasting, whenever k is only modestly greater than d and is much smaller than T .

Nevertheless, the U-MIDAS regression has $k-d$ more parameters under estimation than the constrained MIDAS given by Eq. (6), therefore it would suffer from a substantial loss of degrees of freedom for large k relative to d , and would become even infeasible, whenever the lag order k becomes close to or even greater than the available number of low-frequency observations T . Whereas if the functional constraint was adequate, its ignorance leads to inefficient estimation of parameters for any $k>d$.

3.2 MIDAS

In this section, we discuss the original MIDAS model as proposed by [Ghysels et al. \(2002\)](#) which is given by Eq. (6), also written for convenience as

$$y_t = \beta_0 + \beta_1 \sum_{i=0}^k h(\gamma, i) x_{s(t)-i} + \varepsilon_t, \quad (9)$$

where the constraints on weights $\{w_i\}_{i=0}^k$ in Eq. (6) are explicitly replaced by the restricting function, assuming also that $\sum_{i=0}^k h(\gamma, i) = 1$ holds by construction. Recalling that $d=\dim(\gamma)$, the total number of parameters to be estimated is $d+2$ which corresponds to the elements of vector $\boldsymbol{\theta}=(\beta_0, \beta_1, \gamma')'$. Since model (9) is nonlinear in parameters, the NLS estimator is typically used, yielding

$$\hat{\boldsymbol{\theta}} = \underset{(\beta_0, \beta_1, \gamma')'}{\operatorname{argmin}} \sum_t \left(y_t - \beta_0 - \beta_1 \sum_{i=0}^k h(\gamma, i) x_{s(t)-i} \right)^2 \quad (10)$$

which does not have an explicit solution, but is obtained using numerical optimization.

Assuming again an i.i.d. error term in Eq. (9), it holds under the usual regularity conditions (see, e.g., [Gourieroux and Monfort, 1995](#), chapter 8)

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}_{d+2}, \mathbf{V}_M)$$

as $T \rightarrow \infty$, where $\mathbf{V}_M = \sigma_e^2 \mathbf{S}_M^{-1}$, $\mathbf{S}_M = \operatorname{plim} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{f}_{\boldsymbol{\theta}}^{(1)'} \mathbf{x}_{s(t)}' \mathbf{f}_{\boldsymbol{\theta}}^{(1)} \right)$, where $\mathbf{f}_{\boldsymbol{\theta}}^{(1)} = \partial \mathbf{f}_{\boldsymbol{\theta}} / \partial \boldsymbol{\theta}'$ and

$$\mathbf{f}_{\boldsymbol{\theta}} := (\beta_0, \beta_1 h(\gamma, 0), \beta_1 h(\gamma, 1), \dots, \beta_1 h(\gamma, k))'. \quad (11)$$

It should be pointed out that, apart from the first element of f_{θ} , the other ones are just $\{b_i\}_{i=0}^k$ that were introduced for the U-MIDAS model earlier. Hence, using the exact Taylor expansion of $f_{\hat{\theta}}$ around θ (with some θ_+ “in-between” $\hat{\theta}$ and θ),

$$\sqrt{T}(f_{\hat{\theta}} - f_{\theta}) = \sqrt{T}f_{\theta_+}^{(1)}(\hat{\theta} - \theta) \xrightarrow{d} N\left(\mathbf{0}_{d+2}, f_{\theta}^{(1)} V_M f_{\theta}^{(1)'}\right)$$

as $T \rightarrow \infty$, and one can easily compare the constrained estimates from MIDAS with the unconstrained estimates obtained from U-MIDAS.

This is often of interest on its own, but one can take a step further and, by comparing the statistical significance of the difference between the U-MIDAS and MIDAS estimates, to infer about the adequacy of the functional constraint h . Under the null hypothesis that the MIDAS constraint is nonbinding, the constrained MIDAS and unconstrained U-MIDAS analogous quantities given by β_u and f_{θ} would coincide. Hence, the adequacy of the functional constraint would be linked to the following hypotheses

$$H_0: \beta_u = f_{\theta} \text{ against } H_1: \beta_u \neq f_{\theta}$$

Under H_0 and some usual regularity conditions (see [Kvedaras and Zemlys, 2012](#)) the following holds

$$T(\hat{\beta}_u - f_{\hat{\theta}})A(\hat{\beta}_u - f_{\hat{\theta}}) \xrightarrow{d} \chi^2(k-d),$$

as $T \rightarrow \infty$, where $A = S_x - S_x f_{\theta}^{(1)}(f_{\theta}^{(1)'} S_x f_{\theta}^{(1)})^{-1} f_{\theta}^{(1)'} S_x$. Whereas under the H_1 , the test statistic diverges with probability one.

The NLS estimator is asymptotically efficient for the MIDAS model, whenever model (6) holds with i.i.d. errors and the functional constraint on parameters is adequate. If the adequacy of functional restriction was (strongly) rejected, it cannot be taken for granted that the MIDAS will perform better than the U-MIDAS, i.e., one might be better off to rely on the unconstrained model. However, it is also not obvious that the U-MIDAS will perform better, because the bias connected with the use of an incorrect constraint in the MIDAS regression might be compensated by a substantial reduction of the variance of the NLS estimator, and therefore still improving the efficiency in terms of the mean squared error (MSE) as illustrated, e.g., in [Ghysels et al. \(2016\)](#).

Therefore, given a rejection of the adequacy of the functional constraint, it would be important to understand if the parametric restrictions are sufficient to capture main features of the data generating process (DGP). For instance, the restrictions discussed in [Section 2.1](#) allow for a single hump shape in the weights, are smooth, and impose nonnegativity (no sign switch of $\{w_i\}$), and so on. In some cases it might be that a slight extension, e.g., in the number of parameters, would be sufficient for adequacy, but in other cases a completely different functional form would be needed.

Hence, under the lack of adequacy, it should be worth trying to get some insights by using a nonparametric smoothing of parameters that is considered next.

3.3 MIDAS with nonparametric smoothing of weights

Instead of using a parametric functional constraint, Breitung and Roling (2015) proposed imposing some smoothness on the coefficients $\{b_i\}_{i=0}^k$ in Eq. (7) as measured by their second difference. Namely, define

$$\Delta_2 b_i = \Delta b_i - \Delta b_{i-1} = b_i - 2b_{i-1} + b_{i-2}, \quad i \in \{2, k\},$$

where Δ stands for the difference operator of first order such that $\Delta z_i = z_i - z_{i-1}$ for any real valued quantity z_i indexed by $i \in \mathbb{N}$.

Consider Eq. (7) and let $\beta = (\beta_0, b_0, b_1, \dots, b_k)'$. Furthermore, let the $(k-1) \times (k+2)$ -dimensional differencing matrix

$$\mathbf{D} = \begin{bmatrix} 0 & 1 & -2 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & -2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -2 & 1 & 0 \\ 0 & 0 & \dots & 0 & 1 & -2 & 1 \end{bmatrix},$$

where the first column of zeros is connected with a constant (β_0 in β). Note that

$$\mathbf{D}\beta = (0, \Delta_2 b_0, \Delta_2 b_1, \dots, \Delta_2 b_k)'$$

and hence $\beta' \mathbf{D}' \mathbf{D} \beta = \sum_{i=2}^k (\Delta_2 b_i)^2$.

Given some value $\lambda \in \mathbb{R}_+$ which controls the penalty for nonsmoothness of coefficients $\{b_i\}_{i=0}^k$ in terms of their second differences, the smoothed estimates of β can be obtained as a function of λ using the constrained least squares

$$\hat{\beta}(\lambda) = \underset{\beta_0, b_0, b_1, \dots, b_k}{\operatorname{argmin}} \left\{ \sum_t \left(y_t - \beta_0 - \sum_{i=0}^k b_i x_{s(t)-i} \right)^2 + \lambda \sum_{i=2}^k (\Delta_2 b_i)^2 \right\} \quad (12)$$

$$\begin{aligned} &= \left(\sum_{t=1}^T \mathbf{x}_{s(t)} \mathbf{x}_{s(t)}' + \lambda \mathbf{D}' \mathbf{D} \right)^{-1} \sum_{t=1}^T \mathbf{x}_{s(t)} y_t, \\ &= \left[\mathbf{I}_{k+2} + \frac{\lambda}{T} \mathbf{S}_x^{-1} \mathbf{D}' \mathbf{D} \right]^{-1} \hat{\beta}_u. \end{aligned} \quad (13)$$

Eq. (12) reveals that the estimator is biased whenever $\lambda \neq 0$. Therefore its relative performance in terms of the MSE will depend on whether the induced bias is greater or less than the reduction in variance of the estimator,

similar to the standard MIDAS regression whenever the imposed functional constraint h is incorrect. However, other than for the MIDAS with an inadequate constraint, it also becomes clear from Eq. (13) that, for any fixed λ , the bias vanishes under $T \rightarrow \infty$, as $\hat{\beta}(\lambda)$ approaches the unconstrained $\hat{\beta}_u$, which is unbiased under the linearity of the regression function and efficient among linear estimators whenever the errors of Eq. (7) are i.i.d., although not as efficient as the (nonlinear) MIDAS-based NLS estimator under the adequate constraint.

As can be seen from Eq. (13), $\hat{\beta}(\lambda)$ is a shrinkage estimator and can range from the OLS estimates for $\lambda=0$ to zero under $\lambda \rightarrow \infty$ (not including β_0). Hence it becomes important to select the proper smoothing parameter λ via for example leave-one-out cross-validation, although Breitung and Roling (2015) also propose a selection based on the modified Akaike criterion.

It is clear that not all MIDAS restrictions need to be smooth in the above-defined sense (see, e.g., the step-function restrictions in Ghysels, 2013), but it is clearly useful when it does, especially if one rejects the adequacy of potential functional constraints in the parametric MIDAS framework.

4 Nonlinear parametric MIDAS models

In this section, we will consider models as in Eq. (1), whenever the regression function g is parametric and nonlinear. Hence, the form of both the g and the restriction function h will be assumed to be known, whereas the corresponding parameters in β and γ are not and need to be estimated.

As previously, let us include the functional constraint on parameters explicitly in the model assuming that $\sum_{i=0}^k h(\gamma, i) = 1$ holds by construction, therefore yielding

$$y_t = g\left(\sum_{i=0}^k h(\gamma, i)x_{s(t)-i}; \beta\right) + \varepsilon_t. \quad (14)$$

The i.i.d. error term assumption will be maintained throughout. In addition, g is a regression function and h is an adequate functional constraint.

4.1 General considerations

Let the parameter vector of all coefficients in (14) be denoted by $\theta = (\beta', \gamma')'$. Given that g is a regression function and its parameters are identified in model (14), it can be consistently estimated with the NLS yielding

$$\hat{\theta} = \underset{(\beta', \gamma')'}{\operatorname{argmin}} \sum_t \left(y_t - g\left(\sum_{i=0}^k h(\gamma, i)x_{s(t)-i}; \beta\right) \right)^2, \quad (15)$$

where the solution is obtained using numerical optimization. Under i.i.d. error term in Eq. (14) and usual regularity conditions, the following holds (it is also clear that the results provided for the quasi-linear MIDAS is a special case of this):

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N\left(\mathbf{0}_{\dim(\boldsymbol{\theta})}, \mathbf{V}_N\right) \quad (16)$$

as $T \rightarrow \infty$, where $\mathbf{V}_N = \sigma_e^2 \mathbf{S}_N^{-1}$, $\mathbf{S}_N = \text{plim}\left(\frac{1}{T} \sum_{t=1}^T \frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial g_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}'\right)$, where

$$g_t(\boldsymbol{\theta}) := g\left(\sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(t)-i}; \boldsymbol{\beta}\right). \quad (17)$$

4.2 Logistic smooth transition MIDAS (LSTR-MIDAS)

The logistic smooth transition (LSTR) MIDAS proposed by Galvão (2013) allows for the impact of the high-frequency components to change smoothly with the size of the total high-frequency variable-related quantity. In particular, in the smooth transition MIDAS the general regression function in (17) takes the following form

$$g_t(\boldsymbol{\theta}) = \beta_0 + \beta_1 \sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(t)-i} \left[1 + \beta_2 \mathcal{G}\left(\sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(t)-i}; \beta_3, \beta_4\right) \right], \quad (18)$$

where $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \boldsymbol{\gamma}')' \in \mathbb{A} \subset \mathbb{R}^{5+d}$, and the nonlinearity-inducing logistic function

$$\mathcal{G}\left(\sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(t)-i}; \beta_3, \beta_4\right) = \left[1 + \exp\left(-\beta_3 \frac{\sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(t)-i} - \beta_4}{\sigma_x}\right) \right]^{-1} \in [0, 1], \quad (19)$$

where $\beta_3 > 0$ is imposed for identification. The normalization by $\sigma_x = \text{Var}(x_\tau)^{1/2}$ is not necessary, but is useful for comparability of the β_3 parameters, whenever several explanatory variables with different variability are under consideration. At the same time, the model can be further generalized by, e.g., allowing for different h and/or $\boldsymbol{\gamma}$ in the linear and the nonlinear parts of the regression function and along other extensions known for the usual LSTR models.

From the regression function (18), it is also clear that the parameter β_2 reveals if this kind of nonlinearity improves upon the usual MIDAS model: β_2 would be zero if the nonlinearity were absent. Direct testing of $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$ is however problematic, as under the null hypothesis of linearity ($\beta_2 = 0$), β_3 and β_4 are not identified. Nevertheless, there are a number of potential solutions (see, e.g., a review in Teräsvirta et al., 2010, chapter 5). Whenever $\beta_2 \neq 0$ and the regression function (18) are identified, its parameters can be estimated using the NLS as described in Section 4.1.

Looking closer at the nonlinearity-inducing function in Eq. (19) it becomes obvious that the nonlinearity here emerges because of the *total* high-frequency variable-related quantity $\sum_{i=0}^k h(\gamma, i)x_{s(t)-i}$ in the numerator.

4.3 MIDAS with min–mean–max effects (MMM-MIDAS)

Next we consider an example of a parametric nonlinear model where the nonlinearity is induced by the relative size of high-frequency observation among $\{x_{s(t)-i}\}_{i=0}^k$ and not by their joint aggregate as in the LSTR-MIDAS. Whenever there is a choice of aggregation function of higher-frequency variables (or more generally, a frequency alignment function as in the MIDAS regression), it is often unclear whether some (weighted) average, minimum, maximum, or some combination should be used? In a nonlinear MIDAS framework, one can get the answer to this question using an extended MIDAS aggregation function along the lines proposed in [Kvedaras and Račkauskas \(2010\)](#). For that purpose, let us replace in Eq. (14) the usual MIDAS restriction function h by

$$\tilde{h}\left(\{x_{s(t)-i}\}_{i=0}^k; \beta_2, \gamma, i\right) = \frac{(k+1)\exp(\beta_2 x_{s(t)-i})}{\sum_{j=0}^k \exp(\beta_2 x_{s(t)-j})} \cdot h(\gamma, i) \Bigg|_{\beta_2=0} = h(\gamma, i), \quad (20)$$

leading to the following regression function in (17)

$$g_t(\boldsymbol{\theta}) = \beta_0 + \beta_1 \sum_{i=0}^k \tilde{h}\left(\{x_{s(t)-i}\}_{i=0}^k; \beta_2, \gamma, i\right) x_{s(t)-i} \quad (21)$$

where $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \gamma')'$. Moreover, it should be pointed out that

$$\begin{aligned} & \sum_{i=0}^k \tilde{h}\left(\{x_{s(t)-i}\}_{i=0}^k; \beta_2, \gamma, i\right) x_{s(t)-i} \\ &= \begin{cases} (k+1)w_{i_{\max}} \max(x_{s(t)}, x_{s(t)-1}, \dots, x_{s(t)-k}), & \beta_2 = \infty \\ \sum_{i=0}^k h(\gamma, i)x_{s(t)-i}, & \beta_2 = 0 \\ (k+1)w_{i_{\min}} \min(x_{s(t)}, x_{s(t)-1}, \dots, x_{s(t)-k}), & \beta_2 = -\infty \end{cases} \end{aligned}$$

where for notational brevity (recall Eq. 1) $w_i := h(\gamma, i)$, $i \in \{0, 1, \dots, k\}$, whereas the corresponding weight indices $i_{\max} = \operatorname{argmax}_{i \in \{0, 1, \dots, k\}} x_{s(t)-i}$ and

$$i_{\min} = \operatorname{argmin}_{i \in \{0, 1, \dots, k\}} x_{s(t)-i}.$$

Therefore, besides the MIDAS constraint h which is linear in the observations, the importance of an individual observation in the generalized restriction

function (20) will be nonlinearly magnified or downgraded depending on their relative size and the parameter value of β_2 as well as its sign.

It is also of interest to point out that, taking the first-order Taylor approximation of the regression function (21) around $\beta_2=0$ yields a representation similar to that of the LSTR-MIDAS model in Eq. (18). Namely, letting the moving average $\bar{x}_{s(t)} = \frac{1}{k+1} \sum_{i=0}^k x_{s(t)-i}$,

$$g_t(\boldsymbol{\theta}) \approx \beta_0 + \beta_1 \sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(t)-i} [1 + \beta_2 (x_{s(t)-i} - \bar{x}_{s(t)})] \quad (22)$$

revealing that the influence of a separate component in the MIDAS term will be magnified (downgraded), whenever the particular observation is higher than the moving average and $\beta_2 > 0$ ($\beta_2 < 0$). It is important to point out that, other than the LSTR-MIDAS, both the MIDAS with the MMM effects regression function (21) and its approximation in (22) are identified under the linearity of the effects. Hence, the significance of the MMM nonlinearity can be directly tested exploring $H_0: \beta_2 = 0$ against $H_1: \beta_2 \neq 0$. Letting $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_5, \boldsymbol{\gamma}')$ and using the approximation in (22) one can further merge the smooth transition and MMM-MIDAS into a single model with the regression function

$$\begin{aligned} g_t(\boldsymbol{\theta}) = & \beta_0 + \beta_1 \sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(t)-i} \\ & \times \left[1 + \beta_2 \mathcal{G} \left(\sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(t)-i}; \beta_3, \beta_4 \right) + \beta_5 (x_{s(t)-i} - \bar{x}_{s(t)}) \right] \end{aligned} \quad (23)$$

that allows for both the joint-term and individual observations induced nonlinearity.

Any of the identified models explicated in this section can be estimated with the NLS relying on the results discussed in the previous sections. However, it is worth pointing out that one should always initiate the analysis from the linear MIDAS in order to insure that $\beta_1 \neq 0$ holds in it. Only then one can turn to nonlinear models, as otherwise, i.e., whenever $\beta_1 = 0$ holds, the parameters β_2 and β_5 in the regression function (23) are not identified.

5 Semiparametric MIDAS models

In this section, we extend the usual parametric MIDAS model with some nonparametric function, and therefore presenting semiparametric MIDAS models. A fully nonparametric model will be used only for a stylized presentation of the nonparametric approach, but it is not under consideration for applications.^a

^aAs is typical, because the *curse of dimensionality* problem leads to very slow convergence rates of estimators particularly for larger values of k with a continuous explanatory variable.

For the estimation of the nonparametric function, the local polynomial averaging will be used based on the kernel estimators. For a stylized presentation of the nonparametric estimation problem, consider a regression of a dependent variable $y_t \in \mathbb{R}$ with a single explanatory variable $z_t \in \mathbb{R}$

$$y_t = g(z_t) + \varepsilon_t,$$

where the regression function $g: \mathbb{R} \rightarrow \mathbb{R}$ is three times continuously differentiable and $\varepsilon_t \sim \text{i.i.d. } (0, \sigma_\varepsilon^2)$ is independent of z_t . Assuming that the density function f_Z of the explanatory variable, is also three times continuously differentiable, is positive at $z \in \mathbb{A} \subset \mathbb{R}$, the nonparametric local p th-order polynomial estimator of $g(z)$ is given by

$$\hat{g}_p(z; h) = \arg \min_{a_0} \min_{a_1, a_2, \dots, a_p} \sum_t \left(y_t - \sum_{i=0}^p a_i (z_t - z)^i \right)^2 K_h(z_t, z), \quad (24)$$

where a nonnegative kernel function $K_h(z_t, z) = K\left(\frac{z_t - z}{h}\right)$ satisfies $\int K(v)dv = 1$, $\int vK(v)dv = 0$, and has a positive $\int v^2 K(v)dv$. Here a bandwidth $h \in \mathbb{R}_+$ determines how “local” the averaging will be (how “many” observations around z are taken into account). Whereas the polynomial order p determines the “shape” of the local smoothing: the *local constant* estimator is obtained with $p=0$, $p=1$ yields the *local linear* estimator, and so on.

Maintaining the i.i.d. assumption and under certain regularity conditions (see, e.g., [Li and Racine, 2006](#), chapter 2), including that the bandwidth h satisfies $h \rightarrow 0$, $T \cdot h \rightarrow \infty$, and $T \cdot h^7 \rightarrow 0$,

$$\sqrt{Th} \left(\hat{g}_p(z; h) - g(z) - h^2 b_p(z) \right) \xrightarrow{d} N(0, \kappa \sigma_\varepsilon^2 / f_Z(z)), \quad (25)$$

where the asymptotically vanishing influence of the bias term $b_p(z)$ is a p -specific function of f_Z and g and/or their derivatives, whereas κ is a kernel-specific constant (see [Li and Racine, 2006](#), chapter 2).

The result in (25) remains valid whenever the smoothing parameter h is chosen by the cross-validation (see [Hall et al., 2007](#)). Furthermore, under the general conditions, the cross-validation selected bandwidth h_{CV} satisfies $T^{\frac{1}{5}} h_{CV} \xrightarrow{P} c$, for some positive constant c . Hence, putting aside the issue of the bias, the pointwise convergence in probability of $\hat{g}_p(z; h) - g(z)$ is obtained at the $T^{-\frac{2}{5}}$ rate, whenever there is a single explanatory variable as considered here.

Because of the presence of the bias in Eq. (25), the variability bounds are often reported around (the biased) $\hat{g}(z)$ instead of the confidence bounds around $g(z)$. We follow this practice in the empirical implementation.

5.1 MIDAS with partially (quasi)linear effects (PL-MIDAS)

Next we consider a partially linear (PL) MIDAS model where the quasi-linear MIDAS part is augmented with a nonparametric $g(\cdot)$ in terms of vector $\mathbf{z}_t \in \mathbb{R}^q$, $q \in \mathbb{N}$:

$$y_t = \beta \sum_i h(\boldsymbol{\gamma}, i) x_{s(t)-i} + g(\mathbf{z}_t) + \varepsilon_t, \quad (26)$$

where \mathbf{z}_t can include additional low-frequency variables, a few of $\{x_{s(t)-i}\}_{i=0}^k$ components or their aggregates, e.g., the median. It should be noted that there is no constant in the model (26), as it would not be identified.

Let $\boldsymbol{\theta}_h = (\boldsymbol{\theta}', \mathbf{h}')'$, where the vector of parameters of the quasi-linear part of the model $\boldsymbol{\theta} = (\beta, \boldsymbol{\gamma})'$, and the vector of bandwidths $\mathbf{h} \in \mathbb{R}_+^q$, $q \in \mathbb{N}$. For $q \leq 3$, and given some other regularity conditions (see, e.g., chapter 7 of [Li and Racine, 2006](#)), the \sqrt{T} -consistent estimates of parameters of the quasi-linear part and somewhat more slowly converging estimates of the underlying function g can be obtained jointly solving numerically for

$$\hat{\boldsymbol{\theta}}_h = \underset{(\beta, \boldsymbol{\gamma}, \mathbf{h})'}{\operatorname{argmin}} \sum_t \left(y_t - \beta \sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(t)-i} - \hat{g}_{-t}(\mathbf{z}_j; \mathbf{h}) \right)^2 \quad (27)$$

using the cross-validated kernel estimator of g

$$\hat{g}_{-t}(\mathbf{z}_j; \mathbf{h}) = \frac{\sum_{j=1, j \neq t}^T \left(y_j - \beta \sum_{i=0}^k h(\boldsymbol{\gamma}, i) x_{s(j)-i} \right) K_h(\mathbf{z}_j, \mathbf{z}_t)}{\sum_{j=1, j \neq t}^T K_h(\mathbf{z}_j, \mathbf{z}_t)},$$

where the product kernel

$$K_h(\mathbf{z}_j, \mathbf{z}) = \prod_{i=1}^q K_{h_i}(z_j^{(i)}, z^{(i)})$$

and $K_{h_i}(z_j^{(i)}, z^{(i)})$, $i \in \{1, 2, \dots, \dim(\mathbf{z})\}$ stands for any second-order univariate kernel satisfying the usual conditions defined previously.

Given some regularity conditions and the i.i.d. error term, the result in (16) holds for $\hat{\boldsymbol{\theta}}$ of the parametric part of the PL-MIDAS model with the vector $\boldsymbol{\theta}$ as defined in this section and a properly adjusted explanatory variable. Namely,

$$\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}_{d+1}, \mathbf{V}_P),$$

where $\mathbf{V}_P = \sigma_e^2 S_P^{-1}$, with $S_P = \operatorname{plim} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{f}^{(1)'} \tilde{\mathbf{x}}_{s(t)} \tilde{\mathbf{x}}'_{s(t)} \mathbf{f}^{(1)} \right)$, $\tilde{\mathbf{x}}_{s(t)} = \mathbf{x}_{s(t)} - E(\mathbf{x}_{s(t)} | \mathbf{z}_t)$.

As the convergence of parameters in the parametric PL-MIDAS part is of higher (parametric) rate than that of the nonparametric estimator (24), elements of $\hat{\theta}$ can be used to define

$$\tilde{y}_t := y_t - \hat{\beta} \sum_i h(\hat{\gamma}, i) x_{s(t)-i}.$$

Then, using \tilde{y}_t instead of y_t in the estimator (24), \hat{g}_p is obtained without affecting the limiting distribution as if true parameters of the parametric PL-MIDAS part were used. For the case $q=1$, the result provided in Eq. (25) holds.

5.2 The single index MIDAS model (SI-MIDAS)

The single index (SI) MIDAS model is directly obtained from Eq. (1) by eliminating the parametric part connected with function g , e.g., through $\beta \equiv 1$. As before, let us include the functional constraint on parameters explicitly in the model assuming that $\sum_{i=0}^k h(\gamma, i) = 1$ holds by construction, therefore yielding

$$y_t = g\left(\sum_{i=0}^k h(\gamma, i) x_{s(t)-i}\right) + \varepsilon_t. \quad (28)$$

Note that, in order to ensure identifiability of the regression function g , there is no parameter connected with the total impact of the high-frequency variable (previously denoted by β or β_1 in different models). Recall that here $\dim(\gamma) < k$. If there were no functional constraint reducing the number of parameters, Eq. (28) would be the usual single index semiparametric model. Let $z_t(\gamma; \mathbf{x}_{k,t}) := \sum_{i=0}^k h(\gamma, i) x_{s(t)-i}$, where $\mathbf{x}_{k,t} = (x_{s(t)}, x_{s(t)-1}, \dots, x_{s(t)-k})'$, and

$$G(z_t(\gamma; \mathbf{x}_{k,t})) \equiv E[Y_t | z_t(\gamma; \mathbf{x}_{k,t})] = E[g(z_t(\gamma; \mathbf{x}_{k,t})) | z_t(\gamma; \mathbf{x}_{k,t})].$$

Consider a leave-one-out local constant estimator of G

$$G_{-t}(z_t(\gamma; \mathbf{x}_{k,t}); h) = \frac{\sum_{j=1, j \neq t}^T y_j \cdot K_h(z_j(\gamma; \mathbf{x}_{k,j}), z_t(\gamma; \mathbf{x}_{k,t}))}{\sum_{j=1, j \neq t}^T K_h(z_j(\gamma; \mathbf{x}_{k,j}), z_t(\gamma; \mathbf{x}_{k,t}))}. \quad (29)$$

Then, relying on the approach proposed in Härle et al. (1993), it is straightforward to show that the NLS estimator of parameter vector γ

$$\hat{\gamma} = \operatorname{argmin}_{\gamma} \min_h \sum_t \{y_t - G_{-t}(z_t(\gamma; \mathbf{x}_{k,t}); h)\}^2, \quad (30)$$

obtained by jointly minimizing also in terms of the bandwidth h , satisfies under some regularity conditions and the i.i.d. error term

$$\sqrt{T}(\hat{\gamma} - \gamma) \xrightarrow{d} N\left(\mathbf{0}_{\dim(\gamma)}, V_S\right),$$

as $T \rightarrow \infty$, where $V_S = \sigma_e^2 S_S^{-1}$ and $S_S = \text{plim} \left(\frac{1}{T} \sum_{t=1}^T \left(g_t^{(1)} \right)^2 f_\gamma^{(1)'} \tilde{x}_{s(t)} \tilde{x}_{s(t)}' f_\gamma^{(1)} \right)$.

Here $g_t^{(1)} = \partial g(z_t(\gamma; \mathbf{x}_{k,t})) / \partial z_t(\gamma; \mathbf{x}_{k,t})$, $\tilde{x}_{s(t)} = \mathbf{x}_{s(t)} - E[\mathbf{x}_{s(t)} | z_t(\gamma; \mathbf{x}_{k,t})]$, and $f_\gamma^{(1)} = \partial f_\gamma / \partial \gamma'$ with $f_\gamma := (h(\gamma, 0), h(\gamma, 1), \dots, h(\gamma, k))'$.

Since $\hat{\gamma}$ converges at a parametric rate, which is higher than that of the nonparametric estimators, the same distributional result provided in Eq. (25) applies by defining $z_t := z_t(\hat{\gamma}; \mathbf{x}_{k,t})$.

6 Illustration with simulated data

We illustrate the estimation of the previously defined models by using the simulated data. Sections 6.1 and 6.2 describe the DGP and the estimation of the models using R and its package `midasr`. We concentrate only on nonlinear and semiparametric models, since Ghysels et al. (2016) extensively cover the (quasi)linear MIDAS models with their implementation in R. The statistical inference in this and the next section relies on the limiting distributions characterized in the three previous sections and the inference principles explicated in Section 2.3.

6.1 Data generation

A part of the DGP is common in all the cases, namely:

- the number of high-frequency periods per a low-frequency observation $m = 12$;
- the number of lags of high-frequency variable $k = 24$, and therefore 25 weights $\{w_i\}_{i=0}^k$;
- the beta polynomial function with $\gamma_1 = 2$ and $\gamma_2 = 4$ is used in Eq. (3) to generate the function $h(\cdot)$;
- the high-frequency explanatory variable x_t satisfies $(1 - 0.9L)x_t = \zeta_t \sim N(0, 1)$;
- the i.i.d. error term $\varepsilon_t \sim N(0, 1)$ and independent of ζ_s for any s and t ; and
- the autoregressive part of the dependent process y_t , i.e., the left side of

$$(1 - 0.5L)y_t = g \left(\sum_{i=0}^{24} h(\gamma, i)x_{s(t)-i}; \beta \right) + \xi_t. \quad (31)$$

The right side of Eq. (31) is model specific with

$$\xi_t = \begin{cases} g_z(z_t) + \varepsilon_t, & \text{for PL-MIDAS} \\ \varepsilon_t, & \text{otherwise} \end{cases},$$

where $g_z(z_t) = 0.25z_t^3$, $z_t \sim N(0, 1)$ is the part^b of the PL-MIDAS model to be estimated nonparametrically, and $g(\sum_{i=0}^{24} h(\gamma, i)x_{s(t)-i}; \beta)$ equals:

^bThe scaling 0.25 is used to make the range of g_z similar to that of x_t , if evaluated without extremes.

- as in Eq. (18) with $\beta_1 = 1.5$, $\beta_i = 1$, $i \in \{0, 2, 3, 4\}$, for LSTR-MIDAS;
- as in Eq. (21) with $\beta_1 = 1.5$, $\beta_i = 1$, $i \in \{0, 2\}$, for MMM-MIDAS;
- $\beta \sum_{i=0}^{24} h(\gamma, i)x_{s(t)-i}$, $\beta = 1.5$, for PL-MIDAS; and
- $0.03(\sum_{i=0}^{24} h(\gamma, i)x_{s(t)-i})^3$, for SI-MIDAS.^c

The following R code is used to generate the data of several sizes (250 and 500 low-frequency observations) that later on will be employed as observables for the estimation of models, and additional 200 observations for the initiation (burn-in) of the process in each case that are dropped afterward and unavailable at the estimation stage. We fix the seed of random number generator for purpose of replication.

```
library(midasr)
nnbeta <- function(p, k) nbeta(c(1, p), k)
nbetal <- function(p, k) nbeta(p[c(3, 1:2)], k)
set.seed(1)
lstr_sim <- lapply(c(250, 500), function(n) {
  midas_lstr_sim(n, m = 12, theta = nnbeta(c(2, 4), 24),
    intercept = 1, plstr = c(1.5, 1, 1, 1), ar.x = 0.9,
    ar.y = 0.5, n.start = 200)
})
mmm_sim <- lapply(c(250, 500), function(n) {
  midas_mmm_sim(n, m = 12, theta = nnbeta(c(2, 4), 24),
    intercept = 1, pmmm = c(1.5, 1), ar.x = 0.9, ar.y = 0.5,
    n.start = 200)
})
pl_sim <- lapply(c(250, 500), function(n) {
  midas_pl_sim(n, m = 12, theta = nbetal(c(2, 4, 1.5),
    24), gfun = function(x) 0.25 * x^3, ar.x = 0.9,
    ar.y = 0.5, n.start = 200)
})
si_sim <- lapply(c(250, 500), function(n) {
  midas_si_sim(n, m = 12, theta = nnbeta(c(2, 4), 24),
    gfun = function(x) 0.03 * x^3, ar.x = 0.9, ar.y = 0.5,
    n.start = 200)
})
```

The objects produced by the code return the simulated data needed for estimations. Namely, the low-frequency response variable y , the high-frequency predictor variable x and, in case of PL-MIDAS, the additional low-frequency

^cThe scaling by 0.03 here is used to make the range of the function more similar to that of x_t without extremes. It should be pointed out that $0.03(\cdot)^3$ is estimated nonparametrically, i.e., 0.03 is not a parameter of interest to be estimated—only the λ .

explanatory variable z are generated. y and z have their frequency defined as 1 and x has the frequency 12. These variables are stored as separate elements of each model-associated list component, corresponding to 250 and 500 observations. All variables are objects of class `ts`.

Note that function `nbeta` defined in `midasr` package that implements the normalized *beta polynomial* constraint has the first parameter as a multiplying constant. Hence in the code above the function `nnbeta` is further introduced that omits this parameter, i.e., imposes its value 1. Function `nbeta1` that just changes the order of coefficients of `nnbeta` is introduced purely to simplify the presentation of different models in a single table, as will be apparent shortly.

6.2 Estimation

We now present for each model the estimation code and results. Next, we plot the true weights of the restriction function $\{h(\gamma, i)\}_{i=0}^k$ with their estimated analogs $\{h(\hat{\gamma}, i)\}_{i=0}^k$ and their 95% confidence bounds. In the LSTR-MIDAS case, the true logistic function and its estimate are also reported, whereas in the PL-MIDAS and SI-MIDAS, the true function to be estimated nonparametrically and its estimate are reported instead. These results are presented for several sample sizes.

The following code estimates the models and puts them into the lists, given the respective simulated data with 250 and 500 observations. New function `m1sd` is introduced to deal with lags of high-frequency values. To match the frequencies it uses available date information contained in the objects, in this case standard time series objects of class `ts` in R. Thus instead of passing the frequency ratio in the `m1s` function used for the linear models as described in Ghysels et al. (2016), now simply the low-frequency variable is passed in the `m1sd` function.

The definition of starting values is another difference as compared to the (quasi) linear MIDAS models implemented with `midasr_r` function in Ghysels et al. (2016). Now the supply of starting values is explicitly connected with the specific model. For the LSTR and MMM models, this is done using `midasr_nlpr` function defining in the list of starting values of a high-frequency variable separate coefficients for either “lstr” or “mmm” and the restriction function “r,” which has to be indicated without a scaling constant. For the PL and SI models, the submission of initial values is quite standard.

```
lstr_mod <- lapply(lstr_sim, function(dt) {
  midas_nlpr(y ~ m1sd(x, 0:23, y, nnbeta) + m1sd(y, 1:2,
  y), data = dt, start = list(x = list(lstr = c(1.5,
  1, 1, 1), r = c(2, 4)), y = c(0.5, 0), '(Intercept)' = 1))
})
```

```

mmm_mod <- lapply(mmm_sim, function(dt) {
  midas_nlpr(y ~ mlsd(x, 0:23, y, nnbeta) + mlsd(y, 1:2,
  y), data = dt, start = list(x = list(mmm = c(1.5,
  1), r = c(2, 4)), y = c(0.5, 0), '(Intercept)' = 1))
})

pl_mod <- lapply(pl_sim, function(dt) {
  midas_sp(y ~ mlsd(y, 1:2, y) + mlsd(x, 0:23, y, nbeta1) |
  z, bws = 1, degree = 1, data = dt, start = list(x = c(2,
  4, 1.5), y = c(0.5, 0)))
})

si_mod <- lapply(si_sim, function(dt) {
  midas_sp(y ~ mlsd(y, 1:2, y) | mlsd(x, 0:23, y, nnbeta),
  bws = 1, degree = 1, data = dt, start = list(x = c(2,
  4), y = c(0.5, 0)))
})

```

It is necessary to supply starting values for the model estimation, as is customary for any optimization problem. Note, that NLS problem which each MIDAS estimation needs to solve, is quite sensitive to choice of starting values. However `midasr` package is designed to be used with multiple optimization methods, so various strategies of finding appropriate starting values can be used. The default optimization method for LSTR model is the L-BFGS-B with positive bound on logistic regression term (coefficient $\beta_3 > 0$ in Eqs. 18 and 19), for all the other models the default is Nelder–Mead, with maximum iteration set to 5000. In the simulations with the sample sizes under consideration, the LSTR model seems to be the most awkward.

Besides the correct regression function, we also included the second autoregressive term of the dependent variable in the empirical specification of the estimated models. As it is not present in the DGP, i.e., its parameter is zero, it will be seen that it has indeed—as expected—distinctively larger standard errors and lower significance level.

The estimation results of any specific model can be summarized using the usual `summary` function:

```
summary(lstr_mod[[1]]).
```

Non linear parametric MIDAS regression model with “ts” data:
Start = 3, End = 250.

Formula $y \sim \text{mlsd}(x, 0:23, y, \text{nnbeta}) + \text{mlsd}(y, 1:2, y)$.

Parameters:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|-------------|
| (Intercept) | 1.074729 | 0.119498 | 8.994 | < 2e-16 *** |
| x.lstr1 | 1.643441 | 0.103815 | 15.830 | < 2e-16 *** |

| | | | | |
|---------|-----------|----------|--------|--------------|
| x.lstr2 | 0.750284 | 0.215658 | 3.479 | 0.000598 *** |
| x.lstr3 | 1.786433 | 0.926815 | 1.927 | 0.055103 . |
| x.lstr4 | 1.147379 | 0.727817 | 1.576 | 0.116241 |
| x.r1 | 1.625161 | 0.172832 | 9.403 | < 2e-16 *** |
| x.r2 | 3.279181 | 0.442502 | 7.411 | 2.16e-12 *** |
| y1 | 0.483594 | 0.031593 | 15.307 | < 2e-16 *** |
| y2 | -0.007383 | 0.021986 | -0.336 | 0.737297 |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Residual standard error: 0.9865 on 239 degrees of freedom

which here would summarize the estimated LSTR-MIDAS model with the 250 observations.

The comparison of estimation results of different models is further presented in Table 1 that is produced relying on function `texreg` from the package `texreg` (Leifeld, 2013). The table collects the coefficients together with their standard errors for different models and sample sizes as well as some basic summary statistics.^d This output is obtained using the following code.

```
library(texreg)
texreg(c(lstr_mod, mmm_mod, pl_mod, si_mod), custom.model.names =
  c("LSTR 250",
    "LSTR 500", "MMM 250", "MMM 500", "PL 250", "PL 500",
    "SI 250", "SI 500"), custom.coef.names = c(NA, "slope",
    rep(NA, 7), "slope", NA, "bw", "x.r1", "x.r2", "slope"),
  stars = c(0.001, 0.01, 0.05, 0.1))
```

As far as one can judge from a single realization, the estimation results seem to indicate that overall the sampling properties of the various MIDAS model estimators are quite good and, as expected, tend to improve as the sample size increases. Fig. 1 further illustrates this, namely it characterizes the LSTR-MIDAS model and presents the true MIDAS restriction and the logistic functions used in the DGP together with their estimates. Fig. 2 reports the true MIDAS restriction function from the DGP and its estimate for the MMM-MIDAS model. In Figs. 3 and 4, the true MIDAS restriction and the functions to be estimated nonparametrically in the PL-MIDAS and SI-MIDAS models are plotted together with their estimates.

^dSince the usual R^2 is not apt for nonlinear models, the formula suggested by Hayfield and Racine (2008) is used.

TABLE 1 Estimated models (generated data)

| | LSTR 250 | LSTR 500 | MMM 250 | MMM 500 | PL 250 | PL 500 | SI 250 | SI 500 |
|----------------|-----------------|-----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| (Intercept) | 1.07*** (0.12) | 0.96*** (0.06) | 0.82*** (0.16) | 0.82*** (0.11) | | | | |
| Slope | 1.64*** (0.10) | 1.70*** (0.10) | 1.40*** (0.04) | 1.47*** (0.03) | 1.51*** (0.06) | 1.49*** (0.04) | | |
| x.lstr2 | 0.75*** (0.22) | 0.78*** (0.17) | | | | | | |
| x.lstr3 | 1.79` (0.93) | 1.27** (0.42) | | | | | | |
| x.lstr4 | 1.15 (0.73) | 1.76** (0.59) | | | | | | |
| x.r1 | 1.63*** (0.17) | 1.96*** (0.14) | 2.00*** (0.09) | 2.05*** (0.07) | 2.16*** (0.32) | 2.03*** (0.23) | 3.26*** (0.62) | 2.29*** (0.52) |
| x.r2 | (0.44) | 3.99*** (0.36) | 4.10*** (0.26) | 4.10*** (0.17) | 4.15*** (0.71) | 3.98*** (0.52) | 6.78*** (1.31) | 4.81*** (1.15) |
| y1 | 0.48*** (0.03) | 0.51*** (0.02) | 0.53*** (0.02) | 0.53*** (0.02) | 0.48*** (0.04) | 0.48*** (0.03) | 0.49*** (0.05) | 0.54 (0.04) |
| y2 | -0.01 (0.02) | -0.00 (0.02) | 0.01 (0.02) | -0.02` (0.01) | -0.05` (0.03) | 0.02 (0.02) | 0.06 (0.05) | 0.01 (0.04) |
| x.mmm2 | | | 1.07*** (0.08) | 1.09*** (0.05) | | | | |
| bw | | | | | 0.42* (0.19) | 0.37*** (0.02) | 0.34*** (0.08) | 0.51** (0.19) |
| R ² | 0.97 | 0.97 | 0.95 | 0.96 | 0.94 | 0.92 | 0.54 | 0.58 |
| Num. obs. | 248 | 498 | 248 | 498 | 248 | 498 | 248 | 498 |
| σ ² | 0.99 | 0.95 | 1.02 | 0.98 | 1.05 | 1.14 | 0.98 | 1.02 |

***P<0.001, **P<0.01, *P<0.05, ·P<0.1.

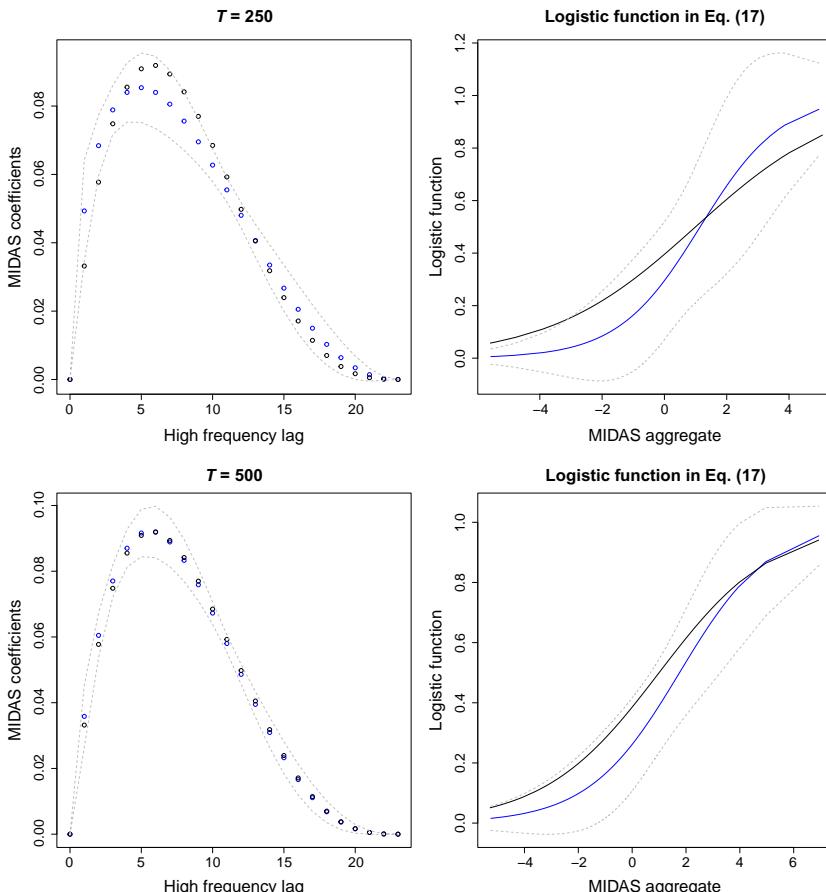


FIG. 1 LSTR-MIDAS: the true and estimated MIDAS restriction function $\{h(\gamma, i)\}_{i=0}^k$ and the logistic function \mathcal{G} as in Eq. (19).

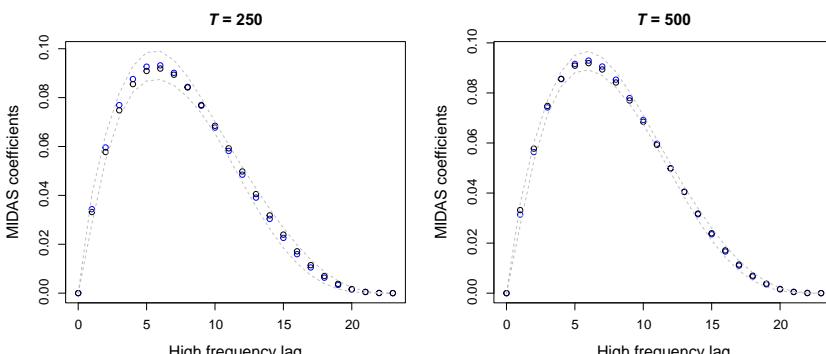


FIG. 2 MMM-MIDAS: the true and estimated MIDAS restriction function $\{h(\gamma, i)\}_{i=0}^k$.

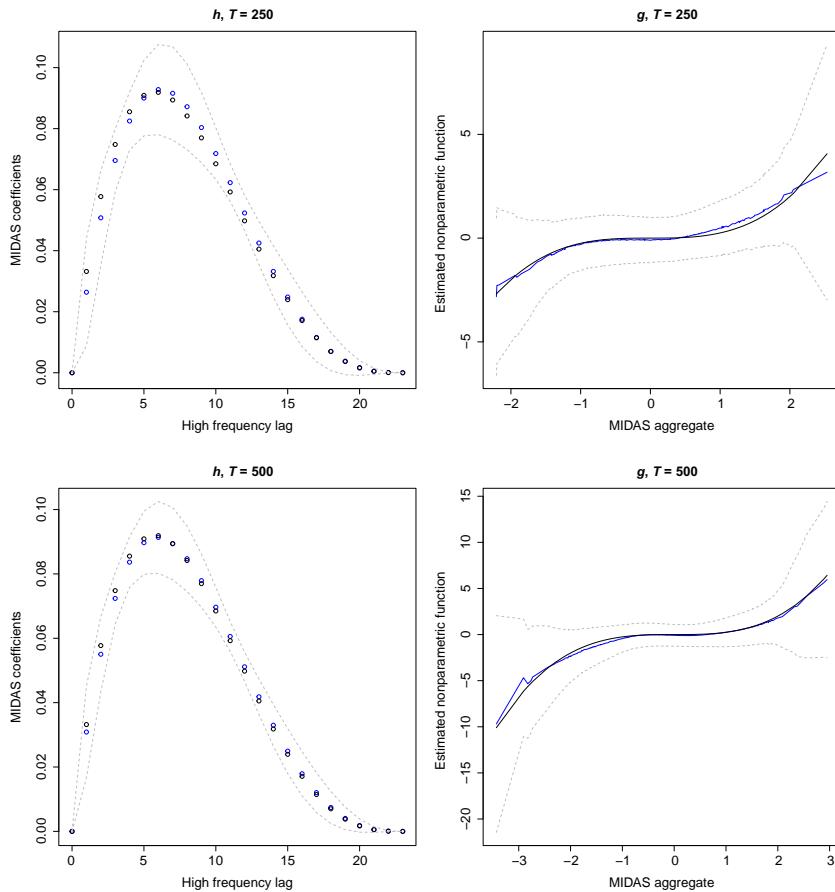


FIG. 3 PL-MIDAS: the true and estimated MIDAS restriction function $\{h(\gamma, i)\}_{i=0}^k$ and the nonparametrically estimated function $g_z(z_t)$.

The estimates of the MIDAS restriction function and the logistic function are reported with their 95% confidence bounds, whereas the 95% variability bounds are plotted around the nonparametric estimates in the PL-MIDAS and SI-MIDAS models.

The function for plotting the model coefficients (and in particular, the implied values of the restriction function) is the same as for parametric MIDAS model. The difference is that for the linear parametric model it made sense to compare the coefficients with U-MIDAS models, for nonlinear and semiparametric models this is not the case. Hence the feature to compare the fitted coefficients with the restriction function used in the model with user supplied parameters was added.

The function to plot the logistic function is a new addition to `midasr` package. The logistic function with the two last LSTR coefficients is

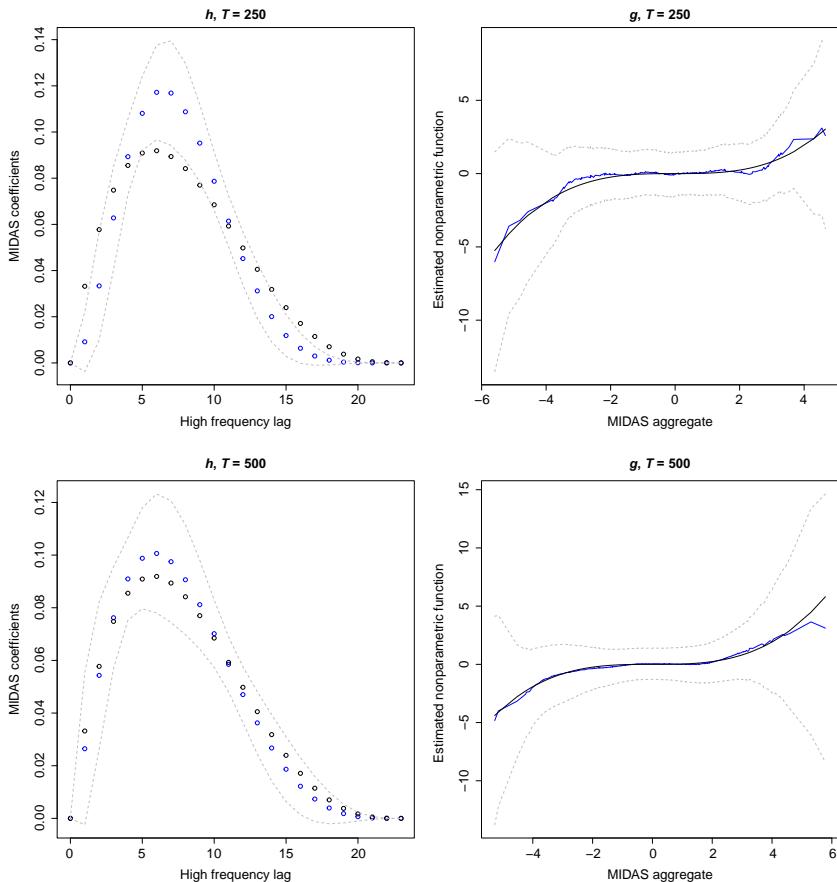


FIG. 4 SI-MIDAS: the true and estimated MIDAS restriction function $\{h(\gamma, i)\}_{i=0}^k$ and the non-parametrically estimated function $g(z_t(\gamma; \mathbf{x}_{k,t}))$.

plotted. Similar to MIDAS coefficient plotting function it is possible to compare the fit with the logistic function calculated with user supplied parameters. The following code produces Fig. 1:

```
par(mfrow = c(2, 2))
plot_midas_coef(lstr_mod[[1]], term_name = "x", compare = c(2,
  4), title = "T = 250")
plot_lstr(lstr_mod[[1]], term_name = "x", compare = list(r = c(2,
  4), lstr = c(1, 1)), title = "Logistic function in eq. (17)")
plot_midas_coef(lstr_mod[[2]], term_name = "x", compare = c(2,
  4), title = "T = 500")
plot_lstr(lstr_mod[[2]], term_name = "x", compare = list(r = c(2,
  4), lstr = c(1, 1)), title = "Logistic function in eq. (17)")
```

Here is the code for producing the Fig. 2.

```
par(mfrow = c(1, 2))
plot_midas_coef(mmm_mod[[1]], term_name = "x", compare = c(2,
  4), title = "T = 250")
plot_midas_coef(mmm_mod[[2]], term_name = "x", compare = c(2,
  4), title = "T = 500")
```

For plotting the estimate of nonparametric function in PL and SI models, function `plot_sp` is used. It only works for univariate nonparametric functions. User can supply her own function for comparison.

In PL-MIDAS the multiplication constant is a part of a specification. To make comparison easier between the models additional feature to set the multiplication constant to 1 of the restriction function was added. If argument `normalize` is set to `TRUE`, simple heuristic is used to find which of the parameters is the multiplication constant and then it is set to 1. Here is the code to produce Fig. 3:

```
par(mfrow = c(2, 2))
plot_midas_coef(pl_mod[[1]], term_name = "x", compare = c(2,
  4), normalize = TRUE, title = "h, T = 250")
plot_sp(pl_mod[[1]], term_name = "z", compare = function(x) 0.25 *
  x^3, title = "g, T = 250")
plot_midas_coef(pl_mod[[2]], term_name = "x", compare = c(2,
  4), normalize = TRUE, title = "h, T = 500")
plot_sp(pl_mod[[2]], term_name = "z", compare = function(x) 0.25 *
  x^3, title = "g, T = 500")
```

Here is how this function is used for producing Fig. 4.

```
par(mfrow = c(2, 2))
plot_midas_coef(si_mod[[1]], term_name = "x", compare = c(2,
  4), title = "h, T = 250")
plot_sp(si_mod[[1]], term_name = "x", compare = function(x) 0.03 *
  x^3, title = "g, T = 250")
plot_midas_coef(si_mod[[2]], term_name = "x", compare = c(2,
  4), title = "h, T = 500")
plot_sp(si_mod[[2]], term_name = "x", compare = function(x) 0.03 *
  x^3, title = "g, T = 500")
```

7 Empirical examples

We provide two examples with macroeconomic data that illustrate the usage of the models covered in the previous sections. In the first case, the mixed frequency data are used to estimate the Okun's law-like equation similar to that considered in [Kvedaras and Račkauskas \(2010\)](#) explaining the yearly growth

rates of gross-domestic product (GDP) by the monthly changes in the unemployment rate. The second example applies the MIDAS-type regression to estimate the impact of the weekly effective federal funds rate on the quarterly consumer prices inflation a quarter ahead.

7.1 Okun's law

We evaluate the potential loss of production caused by an increase in unemployment by relating the growth rates of GDP to the unemployment rate changes in the following way:

$$\Delta \log Y_t = g \left(\sum_{i=0}^k h(\gamma, i) \Delta U_{12t-i} \right) + \zeta_t, \quad (32)$$

where it is expected that the first derivative of g is negative ($\beta_1 < 0$ in parametric specifications) and, due to the constancy of the low-to-high-frequency ratio, i.e., 12 months in a year, it is exploited that $s(t) = 12t$. The Δ stands for the first difference operator, Y and U denote the GDP and unemployment rate, and for the restriction function h the normalized beta polynomial is used.

The models are estimated using yearly US GDP data (in billions of U.S. dollars) of the 1948–2011 period from the U.S. Department of Commerce, Bureau of Economic Analysis (64 observations) and monthly U.S. unemployment rate data (in %) of the same period coming from the U.S. Bureau of Labor Statistics (768 observations) with the sample data available in the `midasr` package as variables `USunempr` and `USrealgdp`.

Although we are not much interested in the quasi-linear model, we start the analysis with it and use its coefficients as starting values for other nonlinear and semiparametric models. Including 2-year monthly lags into the model, which also is supported by the Akaike information criterion, the following regression model therefore serves as a starting point

$$\Delta \log Y_t = \beta_0 + \beta_1 \sum_{i=0}^{23} h(\gamma, i) \Delta U_{12t-i} + \xi_t. \quad (33)$$

Subsequently, the nonlinear and semiparametric models are estimated. It should be pointed out that for the PL-MIDAS, we include the trend series as an explanatory variable, as it was a significant variable in [Kvedaras and Račkauskas \(2010\)](#) capturing the exogenous decrease in the GDP growth rates over the period. Where appropriate, the additional starting values of parameters augmenting those of the quasi-linear MIDAS model are introduced.

The following code produces and summarizes the estimates that are reported in [Table 2](#).

TABLE 2 Estimated models (Okun's law)

| | MIDAS | LSTR | MMM | PL | SI |
|----------------|-----------------|-----------------|-----------------|-----------------|----------------|
| (Intercept) | 0.03*** (0.00) | 0.03*** (0.00) | 0.03*** (0.00) | | |
| x.r1 | 2.29*** (0.34) | 2.35*** (0.46) | 2.09*** (0.38) | 2.31*** (0.37) | 2.28*** (0.42) |
| x.r2 | 3.88*** (0.67) | 3.79*** (0.86) | 3.17*** (0.75) | 3.81*** (0.72) | 3.75*** (0.81) |
| Slope | -0.25*** (0.02) | -0.62** (0.27) | -0.25*** (0.02) | -0.24*** (0.01) | |
| x.lstr2 | | -2.00*** (0.38) | | | |
| x.lstr3 | | 0.13*** (0.01) | | | |
| x.lstr4 | | 1.56** (0.29) | | | |
| x.mmm2 | | | -0.26* (0.13) | | |
| bw | | | | 6.42** (2.80) | 0.14 (0.13) |
| R ² | 0.83 | 0.84 | 0.84 | 0.88 | 0.83 |
| Num. obs. | 62 | 62 | 62 | 62 | 62 |
| σ^2 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |

***P<0.01, **P<0.05, *P<0.1.

```

data(USunemp)
data(USrealgdp)
dt <- list(y = diff(log(USrealgdp)), x = diff(USunemp))
dt$z<-1: length (dt$y)
eqm <- midas_r(y ~ mlsd(x, 0:23, y, nbeta1), data = dt,
  start = list(x = c(2, 4, 0)))
start_lstr <- list(r = coef(eqm, term_names = "x")[1:2],
  lstr = c(coef(eqm, term_names = "x")[3], 1, 1, 0))
lstr_okun <- midas_nlpr(y ~ mlsd(x, 0:23, y, nnbeta), data = dt,
  start = list(x = start_lstr, '(Intercept)' = coef(eqm)[1]),
  method = "Nelder-Mead", control = list(maxit = 5000))
start_mmm <- list(r = coef(eqm, term_names = "x")[1:2],
  mmm = c(coef(eqm, term_names = "x")[3], 0))
mmm_okun <- midas_nlpr(y ~ mlsd(x, 0:23, y, nnbeta), data = dt,
  start = list(x = start_mmm, '(Intercept)' = coef(eqm)[1]))
start_pl <- list(x = setNames(coef(eqm, term_names = "x"),
  NULL))
pl_okun <- midas_sp(y ~ mlsd(x, 0:23, y, nbeta1) | z, data = dt,
  bws = 1, start = start_pl)
start_si <- list(x = setNames(coef(eqm, term_names = "x")[c(1:2)],
  NULL))
si_okun <- midas_sp(y ~ mlsd(x, 0:23, y, nnbeta), data = dt,
  bws = 1, start = start_si)

texreg(list(eqm, lstr_okun, mmm_okun, pl_okun, si_okun),
  custom.model.names = c("MIDAS", "LSTR", "MMM", "PL",
  "SI"), custom.coef.names = c(NA, "x.r1", "x.r2",
  "slope", rep(NA, 2), "slope", rep(NA, 3), "slope",
  NA, "bw"), stars = c(0.01, 0.05, 0.1))

```

Whereas the following code collects the figures produced using the same plotting functions as previously into a single ([Fig. 5](#)).

```

par(mfrow = c(2, 4))
plot_midas_coef(lstr_okun, term_name = "x", title = "LSTR-MIDAS")
plot_midas_coef(mmm_okun, term_name = "x", title = "MMM-MIDAS")
plot_midas_coef(pl_okun, term_name = "x", normalize = TRUE,
  title = "PL-MIDAS")
plot_midas_coef(si_okun, term_name = "x", title = "SI-MIDAS")
plot_lstr(lstr_okun, term_name = "x", title = "Logistic function
  ineq. (17)")
plot.new()
plot_sp(pl_okun, term_name = "z", title = "g(z), T = 62")
plot_sp(si_okun, term_name = "x", title = "g(h(x,gamma)), T=62")

```

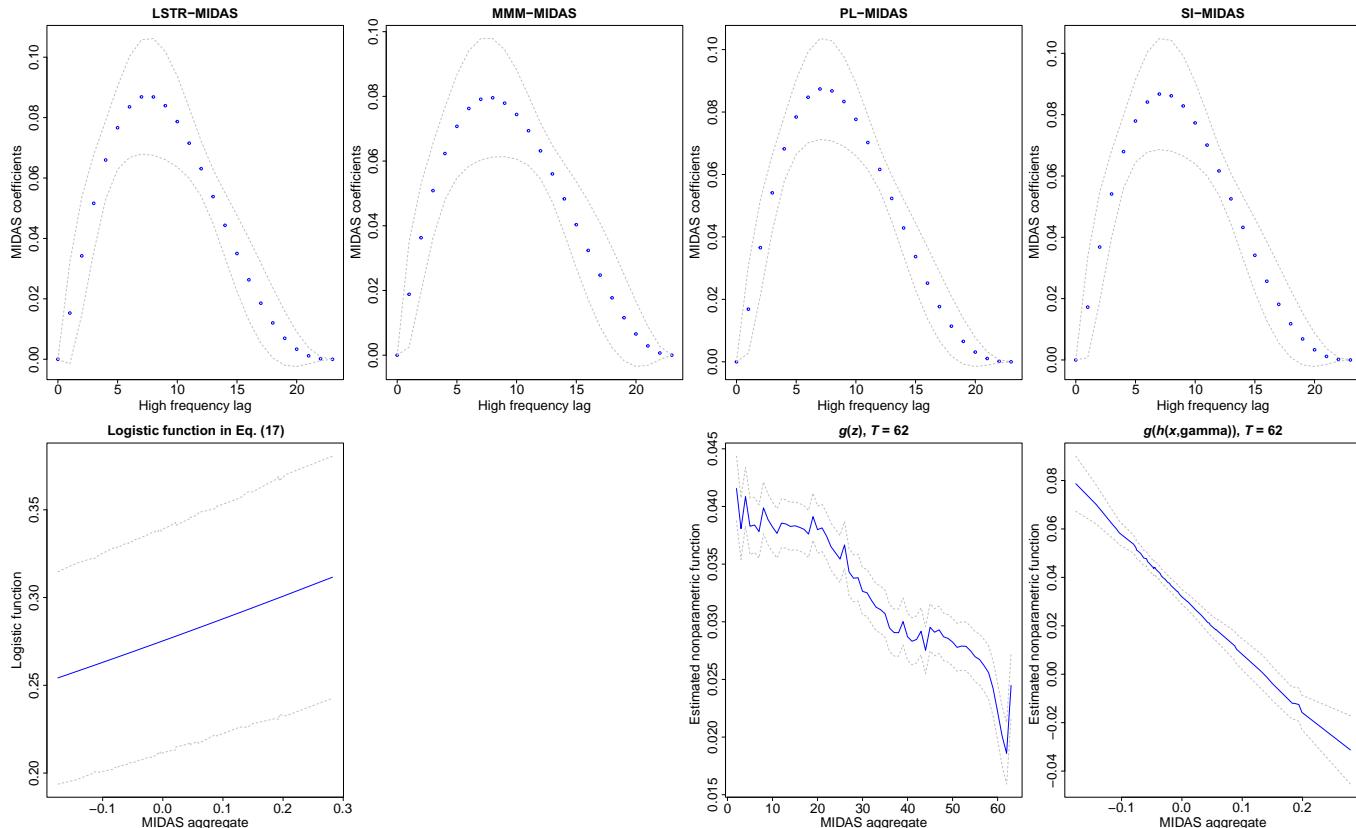


FIG. 5 The Okun empirics: the estimated functions in the LSTR-MIDAS, MMM-MIDAS, PL-MIDAS, and SI-MIDAS models: provided in pairs of figures by columns from left to right, respectively.

Although the nonlinear terms both in the LSTR-MIDAS and MMM-MIDAS models seem to be significant at certain traditional significance levels, they just barely improve (if at all) the precision of the standard quasi-linear MIDAS model and the best fit clearly comes from the PL-MIDAS model which takes into account the exogenous trend of diminishing GDP growth rates that was established to be significant also in [Kvedaras and Račkauskas \(2010\)](#). Otherwise, the usual MIDAS model seems to perform quite well and the SI-MIDAS also does not reveal a substantial nonlinearity as is seen in [Fig. 5](#) producing corresponding plots of estimates of various functions relevant for each model.

In connection with the previously identified best PL-MIDAS model, it is of interest to point out, looking at the PL-MIDAS plots in [Fig. 5](#), that the reduction of the growth rates unconnected with higher unemployment rates initiated approximately from 1970.

7.2 Inflation and the effective federal funds rate

The previous example produced significant nonlinearities apart from the SI-MIDAS. Hence, in this section we concentrate on it and provide an example with a substantial nonlinear effect observed in the SI-MIDAS model. For that purpose we employ a forecasting equation of consumer prices (the consumer price index (CPI) of total items for the United States, P_t) by the means of weekly observations of the effective federal funds rate (r_τ). Again looking at the 2-year period with $k = 103$, we employ the following specification

$$\alpha(L)\Delta\pi_{t+1} = g\left(\sum_{i=0}^{103} h(\gamma, i)\Delta r_{s(t)-i}\right) + \zeta_{t+1}, \quad (34)$$

where the (approximately annualized) quarterly inflation rate $\pi_t = 4(\log P_t - \log P_{t-1})$, with its additional first-order differencing Δ used to remove a potential unit root; the polynomial of autoregressive terms $\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2$ where the lag shift operator is such that $L^j \pi_t = \pi_{t-j}$; and the normalized beta polynomial as defined in [Section 2.1](#) is employed in Eq. (3) to generate the restriction function h .

The quarterly seasonally adjusted data of the CPI of total item for the United States will be used together with the weekly data of effective federal funds rate (in %) both taken from the Federal Reserve Economic Database (FRED). Because the experience of turbulent price developments with stagflation in the 1970s and the spread of importance of rational expectations had substantially changed the conduct of monetary policy, and the recent financial crisis has also led to some particular peculiarities of the monetary policy connected with quantitative easing and the zero lower bound problems, we restrict ourself to the relatively normal 1980Q1-2007Q4 period (with initial 112 quarterly and 1448 weekly observations in it) in order to avoid additional modeling of structural breaks.

The time series structure of the data is represented by `data.frame`, one column containing the year, month, and day indicating the start of the week or the quarter. We use R library `xts` to convert the data into `xts` objects and then to make the necessary transformations. Note that since we are modeling one quarter into the future, we use the negative lag to shift the time series upward. The following code prepares the data.

```
library(xts)
data(UScpiqs)
data(USEffrw)
P <- xts(UScpiqs[, 2], order.by = as.Date(UScpiqs[, 1]))
r <- xts(USEffrw[, 2], order.by = as.Date(USEffrw[, 1]))
dpit <- diff(4 * diff(log(P)))
dr <- diff(r)
data_cpi <- list(y = lag(dpit["1980-01-01/2007-10-01"],
k = -1L, na.pad = FALSE), x = dr)
```

It should be further noted that the effective data sample becomes shorter because of one-step-ahead forecasting and the presence of two lags of the dependent variable (as well as about 2-year lags of the high-frequency variable).

As previously, our modeling strategy is to start from a quasi-linear MIDAS model and use its estimated coefficients that are relevant for the SI-MIDAS as the starting values. The model is specified again with function `m1sd`, which uses the date information in the `xts` objects to align the data. The code that follows performs these estimations.

```
midas_cpi <- midas_r(y ~ m1sd(y, 1:2, y) + m1sd(x, 0:103,
y, nbeta1), data = data_cpi, start = list(x = c(3, 3,
0)))
start_si <- list(x = coef(midas_cpi, term_names = "x")[1:2],
y = coef(midas_cpi, term_names = "y"))
start_si <- lapply(start_si, setNames, NULL)
si_cpi <- midas_sp(y ~ m1sd(y, 1:2, y) | m1sd(x, 0:103,
y, nnbeta), bws = 1, data = data_cpi, start = start_si)
```

The results are presented in [Table 3](#), whereas the connected graphs are in [Fig. 6](#).

There is a substantial asymmetry in the impact of the increasing/decreasing effective federal funds rate. Furthermore, a significant inflation-reducing impact appears only whenever the increase of the rates is sufficiently large. Given such a nonlinearity, it is not unexpected that the estimates of the quasi-linear MIDAS model were insignificant at the usual significance levels, whereas the parameters of the SI-MIDAS are highly significant. However, it seems to have only minor effect on the predictive precision (the R^2 increases from 0.3602 to 0.3632).

TABLE 3 Estimated models (CPI)

| | MIDAS | SI |
|----------------|-----------------|-----------------|
| (Intercept) | -0.00 (0.00) | |
| y1 | -0.55*** (0.11) | -0.56*** (0.09) |
| y2 | -0.46*** (0.12) | -0.47*** (0.08) |
| x1 | 4.03 (6.16) | 4.11** (1.62) |
| x2 | 4.98 (6.29) | 4.93** (2.02) |
| x3 | -0.07 (0.08) | |
| bw | | 0.04* (0.02) |
| R ² | 0.36 | 0.36 |
| Num. obs. | 109 | 109 |
| σ ² | 0.02 | 0.02 |

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.

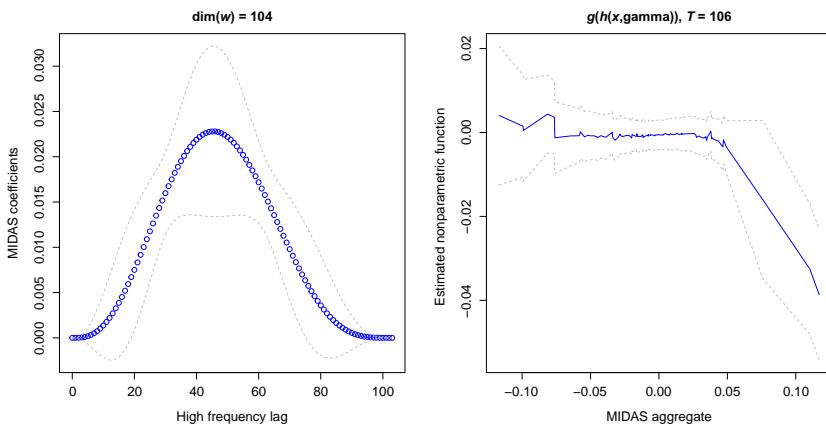


FIG. 6 The CPI empirics: the estimated functions in the SI-MIDAS model.

References

- Andreou, E., Ghysels, E., Kourtellos, A., 2011. Forecasting with mixed-frequency data. In: Clements, M., Hendry, D. (Eds.), Oxford Handbook of Economic Forecasting. Oxford University Press, Oxford, pp. 225–245. <http://www.oxfordhandbooks.com/view/>. <https://doi.org/10.1093/oxfordhb/9780195398649.001.0001/oxfordhb-9780195398649-e-9>.
- Andreou, E., Ghysels, E., Kourtellos, A., 2013. Should macroeconomic forecasters use daily financial data and how? J. Bus. Econ. Stat. 31 (2), 240–251. <https://doi.org/10.1080/07350015.2013.767199>.

- Armesto, M.T., Engemann, K.M., Owyang, M.T., 2010. Forecasting with mixed frequencies. *Fed. Reserve Bank St. Louis Rev* 92, 521–536.
- Bai, J., Ghysels, E., Wright, J.H., 2013. State space models and MIDAS regressions. *Econ. Rev.* 32 (7), 779–813. <https://doi.org/10.1080/07474938.2012.690675>.
- Breitung, J., Roling, C., 2015. Forecasting inflation rates using daily data: a nonparametric MIDAS approach. *J. Forecast.* 34 (7), 588–603. <https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2361>.
- Diebold, F.X., Mariano, R.S., 2002. Comparing predictive accuracy. *J. Bus. Econ. Stat.* 20 (1), 134–144. <https://doi.org/10.1198/073500102753410444>.
- Foroni, C., Marcellino, M.G., 2013. A survey of econometric methods for mixed-frequency data. SSRN scholarly paper ID 2268912, Social Science Research Network, Rochester, NY. <https://papers.ssrn.com/abstract=2268912>.
- Foroni, C., Marcellino, M., Schumacher, C., 2015. Unrestricted mixed data sampling (U-MIDAS): MIDAS regressions with unrestricted lag polynomials. *J. R. Stat. Soc. A. Stat. Soc.* 57–82. <https://doi.org/10.1111/rssa.12043>.
- Foroni, C., Marcellino, M.G., Stevanovi, D., 2018. Mixed frequency models with MA components. SSRN scholarly paper ID 3127429, Social Science Research Network, Rochester, NY. <https://papers.ssrn.com/abstract=3127429>.
- Galvão, A.B., 2013. Changes in predictive ability with mixed frequency data. *Int. J. Forecast.* 29 (3), 395–410. <http://www.sciencedirect.com/science/article/pii/S0169207012001689>.
- Ghysels, E., 2013. MATLAB toolbox for mixed sampling frequency data analysis using MIDAS regression models. Available on MATLAB Central at [http://www.mathworks.com/matlabcentral/fileexchange/45150](http://www.mathworks.com/matlabcentral/fileexchange/45150-midas-regression), <https://www.mathworks.com/matlabcentral/fileexchange/45150>.
- Ghysels, E., 2016. Macroeconomics and the reality of mixed frequency data. *J. Econ.* 193 (2), 294–314. <http://www.sciencedirect.com/science/article/pii/S0304407616300653>.
- Ghysels, E., Marcellino, M., 2018. *Applied Economic Forecasting Using Time Series Methods*. Oxford University Press, Oxford, New York.
- Ghysels, E., Qian, H., 2019. Estimating MIDAS regressions via OLS with polynomial parameter profiling. *Economet. Stat.* 9, 1–16. <http://www.sciencedirect.com/science/article/pii/S2452306218300066>.
- Ghysels, E., Valkanov, R., 2012. Forecasting volatility with MIDAS. In: *Handbook of Volatility Models and Their Applications*. Wiley-Blackwell, pp. 383–401. <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118272039.ch16>.
- Ghysels, E., Santa-Clara, P., Valkanov, R., 2002. The MIDAS touch: mixed data sampling regression models. Working paper, UNC and UCLA.
- Ghysels, E., Santa-Clara, P., Valkanov, R., 2006. Predicting volatility: getting the most out of return data sampled at different frequencies. *J. Econ.* 131 (1), 59–95. <http://www.sciencedirect.com/science/article/pii/S0304407605000060>.
- Ghysels, E., Sinko, A., Valkanov, R., 2007. MIDAS regressions: further results and new directions. *Econ. Rev.* 26 (1), 53–90. <https://doi.org/10.1080/07474930600972467>.
- Ghysels, E., Kvedaras, V., Zemlys, V., 2016. Mixed frequency data sampling regression models: the *R* package midasr. *J. Stat. Softw.* 72(4). <http://www.jstatsoft.org/v72/i04/>.
- Gourioux, C., Monfort, A., 1995. *Statistics and Econometric Models*. vol. 1. Cambridge University Press.
- Hall, P., Li, Q., Racine, J.S., 2007. Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Rev. Econ. Stat.* 89 (4), 784–789. <https://www.jstor.org/stable/40043100>.

- Hardle, W., Hall, P., Ichimura, H., 1993. Optimal smoothing in single-index models. *Ann. Stat.* 21 (1), 157–178. <https://projecteuclid.org/euclid-aos/1176349020>.
- Hayfield, T., Racine, J., 2008. Nonparametric econometrics: the np package. *J. Stat. Softw.* 27(5). <https://www.jstatsoft.org/article/view/v027i05>.
- Kvedaras, V., Račkauskas, A., 2010. Regression models with variables of different frequencies: the case of a fixed frequency ratio. *Oxf. Bull. Econ. Stat.* 72 (5), 600–620. <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0084.2010.00585.x>.
- Kvedaras, V., Zemlys, V., 2012. Testing the functional constraints on parameters in regressions with variables of different frequency. *Econ. Lett.* 116 (2), 250–254. <http://www.sciencedirect.com/science/article/pii/S0165176512000961>.
- Leifeld, P., 2013. texreg: conversion of statistical model output in R to LATEX and HTML tables. *J. Stat. Softw.* 55(8). <https://www.jstatsoft.org/article/view/v055i08>.
- Li, Q., Racine, J.S., 2006. Nonparametric Econometrics: Theory and Practice. Princeton University Press. <https://ideas.repec.org/b/pup/pbooks/8355.html>.
- Rodriguez, A., Puggioni, G., 2010. Mixed frequency models: Bayesian approaches to estimation and prediction. *Int. J. Forensic Dent.* 26 (2), 293–311. <http://www.sciencedirect.com/science/article/pii/S0169207010000154>.
- Teräsvirta, T., Tjøstheim, D., Granger, C.W.J., 2010. Modeling Nonlinear Economic Time Series. Oxford University Press. <http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199587148.001.0001/acprof-9780199587148>.
- White, H., 1981. Consequences and detection of misspecified nonlinear regression models. *J. Am. Stat. Assoc.* 76 (374), 419–433. <https://www.jstor.org/stable/2287845>.

Chapter 5

Encouraging private corporate investment in India

Hrishikesh Vinod^{a,*}, Honey Karun^b and Lekha S. Chakraborty^{c,d}

^a*Fordham University, New York, NY, United States*

^b*Senior Resident Representative Office, International Monetary Fund, New Delhi, India*

^c*National Institute of Public Finance and Policy, New Delhi, India*

^d*Levy Economics Institute of Bard College, New York, NY, United States*

*Corresponding author: e-mail: vinod@fordham.edu

Abstract

A typical macroeconomic regression of private corporate investment on public infrastructure investment, rate of interest, private credit, capital flows and output gap involves a mixture of stationary and nonstationary variables. Moreover, the available data series (2011–16) for estimating the regression for India is too short for asymptotic statistical inference. Hence we use maximum entropy (ME) bootstrap from R package “meboot” to confirm positive role of public infrastructure investment. The significant result has policy implications in terms of the current debate whether public investment “crowds-in” rather than “crowds-out” private corporate investment in India. We use another R package “generalCorr” to study whether the right-hand side variables “approximately cause” private investment, or are subject to the endogeneity problem. While finding evidence supporting public infrastructure spending to encourage private investment in India, we highlight new R tools for estimation and inference in many macroeconomic regressions.

JEL Classification: E62, C32, H62

Keywords: Private investment, Time series, Bootstrap, Causality, Crowding out, Real interest rate, Fiscal policy, Monetary policy, Infrastructure investment

1 Introduction

Private investment in India has averaged around 25% of gross-domestic product (GDP) during 2004–05 to 2015–16, wherein both corporate and household sectors consistently contributed more than 10%. Corporate sector contribution during this period peaked at 17.3% of GDP in 2007–08 and declined to 10.6%

in 2011–12 with revival to 15% in 2015–16. Public sector contributed an average of 8%–8.5% of GDP during the same period.

Another noteworthy feature of the Indian economy regarding the investment climate in India was a secular rise in the stalled projects since 2007. Interestingly most of these stalled projects lie in the infrastructure and linked sectors. “The Economic Survey 2014–15 of India” highlighted that the size of stalled projects (as % of GDP) in the private sector was two-to-three times larger than in the public sector.^a

Successive economic surveys in India (for instance, 2012–13, 2013–14, and 2014–15) have highlighted these concerns and pointed to several factors causing the decline in private investment over time. While some of the factors were domestic, such as a change in monetary policy stance to stabilize rising inflation and the household expectations about inflation; others were external spill-over effects of slowing global demand and weak recovery in global growth.

The Economic Survey 2013–14 stressed the severity of challenges in financing private investment and argued that high and persistent inflation along with lower real interest rates are reducing private savings, thus reducing the supply of funds. Accordingly, the survey urged policy measures aimed at reducing the fiscal burden (through fiscal consolidation), stabilizing inflation, and reduction in resource preemption, thereby allowing more financial space for private investment (or reduced “crowding out”). The survey in 2014–15, noted “... the balance sheet syndrome with Indian characteristics creates a web of different challenges that cause hold back private investment...primary investment must remain the primary engine of long-run growth...public investment will have an important role to play....”

These surveys underscored, therefore, the need for reviving complementary public investment in the short run to “crowd-in” private investment. The discussion above shows how successive economic surveys in India approached the issue of decline in private investment and emphasized potential policy levers that may change the course of private investment in India. The survey narrative also highlights that the policy levers designed to encourage private investment in informed policy making process will depend on whether the public investment is argued to be “crowding in” or “crowding out” private investment.

India has recently experienced a private investment slowdown and therefore that understanding the determinants of private corporate investment during this period is an important question for policymakers who want to revive investment. However, the introduction of new national accounts series with revised methodology posed challenges with only a limited number of observations available, and this makes it hard to perform meaningful time series analysis. Our chapter offers a time series technique to overcome some of the

^aSee table 4.1, chapter 4, vol. 1, Economic Survey 2014–15 which can be accessed at: <http://www.indiabudget.gov.in/budget2015-2016/es2014-15/echapvol1-04.pdf>.

statistical challenges that the availability of only a short data series pose. Our chapter uses ME (maximum entropy) bootstrap method to overcome the econometric constraints of using a short time series after the publication of new macroeconomic series in India. Our results reinforce the crowding in properties of public investment in India. They are tested within the context of an econometric model investigating broader determinants of private investment.

The chapter is divided into following sections. [Section 2](#) provides a brief literature review. [Section 3](#) interprets the data and discusses some stylized facts regarding the data on investment in Indian economy. [Section 4](#) explains the methodology and reports our estimated results. Data details are in [Section 5](#). [Section 6](#) provides results from our study of causal paths. [Section 7](#) discusses the implications and concludes.

2 Literature review

The longstanding debate on crowding out effects on private investment—real and financial—is not expected to reach a definitive conclusion in the near future. Blinder and Solow (1973) and Buiter (1990) provided theoretical understanding on crowding out effects while many others tested the theoretical foundations empirically. Recently, in the aftermath of the global crisis, interest in the role of public investment in crowding in (faltering) private investment intensified. The October 2014 World Economic Outlook of International Monetary Fund (IMF) contains a chapter on the macroeconomic effects of public infrastructure investment that does not find a significant effect on private investment for advanced economies. However, there is no recent empirical evidence on crowding out in the context of emerging economies to the best of our knowledge. The results from selected literature are summarized in [Table 1](#). It is clear that the literature provides few definitive results regarding financial crowding out of private investment.

In Indian context, [Chakraborty \(2007\)](#) attempted to explore both real and financial aspects of the crowding out argument and found no evidence for either. In recent years, [Bahal et al. \(2015\)](#) studied the relationship between public capital accumulation and private investment in India. Their chapter observed crowding out effects on private investment during 1950–2012, whereas the opposite results were highlighted for post-1980 period. Their results using quarterly data report crowding in from 1996Q2 onward. [Dash \(2016\)](#) analyzed the impact of public investment on private investment for the period 1970–2013, and found evidence for crowding out, which was subdued during the postliberalization period. Dash, however, did find a positive impact of public infrastructure investment on private investment in the short run.

[Mallick \(2016\)](#) found evidence for the crowding out effect of government investment during 1970–2013 attributing it mostly to noninfrastructure government investment. The study also reported a larger impact of private

TABLE 1 Broad understanding of the crowding out effects on private investment in the literature

| Argument | Literature | Economic rationale |
|---|--|---|
| Real crowding out (in which) | | |
| Public investment crowds out private investment | Blejer and Khan (1984), Cebula (1978), Shafik (1992), Parker (1995), Ostrosky (1979), Tun and Wong (1982), Sundararajan and Thakur (1980), Pradhan et al. (1990), Alesina et al. (2002), and many others | Public investment may substitute private investment when private sector does not utilize public capital for capacity expansion. On the other hand, capital formation by private sector may allow public sector to withhold their investment decisions |
| Real crowding in (in which) | | |
| Public and private investment complement each other | Blejer and Khan (1984), Cebula (1978), Ramirez (1994), Greene and Villanueva (1990), Buiter (1977), Erenburg (1993), Aschauer (1989), and many others | Public investment may complement private when the increase in public capital formation may impact aggregate demand thereby can provide stimulus to private investment |

Source: Authors' analysis.

investment on income than on public investment. Finally, Chhibber and Kalloor (2016) analyzed the determinants of aggregate, as well as, sector-wise (corporate and noncorporate) private investment for the period 1980–81 to 2013–14 and argued for crowding in effects of public investment on private investment.

The empirical literature reviewed here relies on two set of econometric testing methods: autoregressive distributed lag (ARDL) models and vector autoregressive (VAR) models. Both are useful for time series estimations involving longer series, in testing for structural breaks using different specifications to estimate long-term relationships. However, ARDL and VAR models often involve differencing or detrending of variables to deal with the problems associated with ubiquitous nonstationarity of underlying macroeconomic time series. Moreover, these models often yield insignificant results when the time series is short. However, the insignificance might be a statistical artifact preventing a recognition of useful economic signals relevant for short run policy purposes.

Statistical inference involving short and evolving economic time series appears to be excessively focused on the violation of stationarity (also described as the integrated of order zero, or $I(0)$ assumption). A policy-focused macroeconomist is generally faced with models involving a mixture of $I(0)$ and nonstationary $I(d)$ series, where the order of integration “ d ” can be different for different series. If the data have a finite but relatively “long-memory,” then “ d ” is a fraction.

The usual practice of differencing or detrending the series can be inappropriate for mixed $I(0)$ and $I(d)$ data, especially if d is fractional (long-memory) and/or the series is subject to finite structural changes. [Vinod \(2006\)](#) shows that random walk series $I(1)$ arise from an assumption of infinite memory, whereas macroeconomic literature has considerable evidence confirming that economic agents have short memory. Infinite memory implicit in $I(1)$ formulation rules out infrequent jumps often present due to structural changes in legislation or political upheavals or realignments. One can formally test for structural change. However, upon knowing the presence of structural changes it is not clear how to transform the observed series into $I(0)$, needed for the usual statistical inference.

Econometricians have tried to develop approaches to address problems associated with nonstationary data. For instance, vector autoregressions (VAR) involve differencing the nonstationary data to achieve stationarity. However differencing is valid only if the data have unit roots, which in turn requires pretesting for unit roots. It is well known that when any pretesting is done, the size and power of subsequent statistical inference are affected. [Toda and Yamamoto \(1995\)](#) attempted to eliminate such problems. In this chapter, we consider maximum entropy bootstrap (meboot) based on [Efron \(1979\)](#) bootstrap for exploring determinants of private investment in India.

The “meboot” algorithm is a seven-step procedure which allows one to generate replicates or “reincarnations” of the original series, as termed by [Vinod \(2004\)](#), to be used for statistical inferences. The meboot resamples allow overcoming the unit root and structural change pretest problems, while avoiding any differencing-type transformations of original time series simply for ensuring the stationarity assumption.^b Meboot constructs segments of ME density $f(x)$ subjected to certain mass and mean preserving constraints. The maximized entropy is defined in terms of Shannon information in a density $f(x)$ function. The entropy H is defined as: $H = E(-\log f(x))$.

The meboot ensembles or the replicates created from entropy maximizing densities can be proved to satisfy the ergodic theorem, central limit theorem, and Doob’s theorem. In addition, the constructed ensembles have the property of retaining the overall shapes of autocorrelation and partial autocorrelation functions of the original time series data, without imposing parametric constraints.

^bFor further detailed on the seven-step algorithm, see [Vinod \(2006, 2009, 2013\)](#).

Thus, meboot appears to be a useful option for mixed models involving short and nonstationary-dependent data. Fig. 1 shows the actual data on private investment and a sample of three replicas generated from the meboot algorithm. It shows that the basic shape of the nonstationary $I(1)$ series is retained in each replica. Hence we can construct a thousand such replicas to approximate the unknown “population” of analogous time series, eventually allowing construction of confidence intervals for parameters of interest.

The key advantage of meboot is that features of the population are approximated without artificially differencing, detrending, or eliminating structural changes in the series to force them to behave like $I(0)$ series. An appendix provides additional graphs illustrating shape retention by meboot resamples for other variables in our model. Fig. 1 clearly indicates that the meboot resample retains the shape of the original time series under consideration as the resamples are strongly dependent on it. In addition to the meboot confidence intervals based on [Vinod and López-de-Lacalle \(2009\)](#), we also report the asymptotic confidence intervals from OLS regressions for comparison in Tables 5–10.

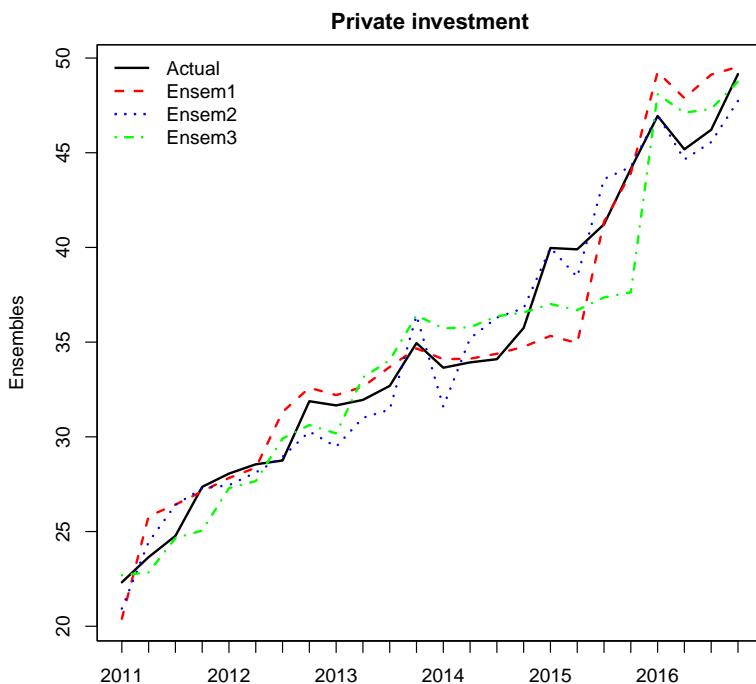


FIG. 1 Actual and generated ensembles of private investment (in INR billion) used for confidence intervals.

3 Interpreting data and model implications

We explore the determinants of private investment following Chakraborty (2007) by incorporating interest rates (both short and long term) in the model equation as below to gage the impact of interest rates on private corporate investment^c:

$$I_{pvt} = a + \beta_1 I_{pub} + \beta_2 i_r + \beta_3 C_{pvt} + \beta_4 K_{forgn} + \beta_5 Y^* + e_t \quad (1)$$

where I_{pvt} =private investment, I_{pub} =public investment, i_r =real interest rate (using two versions: short- or long-term rate), C_{pvt} =credit to the private sector, K_{forgn} =foreign investment capital flows, and Y^* =output gap.

Both the price and quantity of credit variables are added in the model to test McKinnon hypothesis whether cost of the credit or quantity matters for private investment. The quarterly data on macrovariables are sourced from the database of the Reserve Bank of India (RBI) for the period 2011–16. The Central Statistics Office (CSO) of India introduced a new series of national accounts, with certain revisions in the methodology for estimating gross value added and GDP which provides data at 2011–12 prices.^d

Capital formation: The latest available data for detailed information on sector-wise capital formation can be sourced from National Accounts Statistics till 2016–17 at the new base year. We categorized the public investment into infrastructure and noninfrastructure as suggested in Parker (1995).

Investment: The sector-wise data are available on annual basis only. Bahal et al. (2015) has estimated quarterly data by aggregating project-level costs into quarterly time series for sector and industry-level investment activity using CapEx database of the Centre for Monitoring of Indian Economy (CMIE). We estimated the sector-wise quarterly data on investment with certain simplified assumptions, i.e., maintaining the annual relative shares of private corporate and public investment in each quarter. We follow the same assumption to estimate public infrastructure and noninfrastructure data for quarterly series.

Interest rates: Selecting appropriate interest rate from the available spectrum of interest rates in India needed a careful examination of certain stylized facts regarding such rates stated in Table 2.^e

We consider the 91-day treasury bills rate for short-term interest rates. The corresponding short-term rate is shown as solid lines in top right and bottom left panels of Fig. 2. Our long-term interest is 10-year yield on government securities shown as dotted lines in top right and bottom left panels of Fig. 2. We study the impact of these real interest rates on private investment. We subtract the two

^cSee Chakraborty (2007) for detailed derivation of the equation.

^dhttp://www.mospi.gov.in/sites/default/files/press_release/nad_press_release_30jan15.pdf.

^eSee Vinod et al. (2014) and Chakraborty (2007, 2012).

TABLE 2 Stylized facts of major interest rates in India

| Interest rates | Stylized facts |
|---|---|
| Call money market rate | Usually exhibits large volatility |
| Bank rate | Usually nonvarying in nature |
| Prime lending rate | A potential indicator of long-term rate and exhibits stickiness |
| Redemption yield on Government securities | A potential indicator of long-term rate in case of shift from seigniorage financing to bond financing of fiscal deficit |
| Treasury bills rate | A significant reference interest rate in the short term |

Source: Authors' analysis.

inflation rates from nominal interest rates to represent our “real” interest rate series. We have data on two distinct inflation rates showing that the choice matters.

The top left corner of Fig. 2 shows movements of the two inflation rates. The retail inflation based on consumer price index (CPI) is seen to be quite distinct from wholesale inflation based on the wholesale price index (WPI-core). The retail price inflation has remained high and positive while wholesale price inflation has been declining and reached a negative zone in 2015. The movement of interest rates shows that real interest rates have been negative and rising in half of the quarters of the period covered here (right panel above). The bottom left panel of Fig. 2 shows real interest rates based on core WPI-based inflation. Although the real interest rates have been rising for the entire period of our study, the upward movement in “real” interest rates based on WPI (bottom left of Fig. 2) is more steady compared to the less consistently upward trend when based on CPI (top right of Fig. 2). In the interest of brevity, we report estimates using real interest rates based on CPI only.

Output gap: OG is considered to be the most debated macroeconomic variable in the present context. A simple definition of the output gap is:

$$Y^* = OG = [(Actual GDP - Potential GDP)/Potential GDP] \times 100.$$

The potential level of output, though, can be higher than the actual, if the resource utilization is maximized at the potential level; Tanzi (1985) argued that cyclical factors can make the actual output to be below or above the potential output. Nelson and Plosser (1982) state that potential output is often termed as the permanent component of the economy and is usually nonstationary. Thus, the output gap refers to the transitory part which is composed of two unobserved components—cyclical and irregular components. Hence OG is usually assumed to be stationary, or $I(0)$.

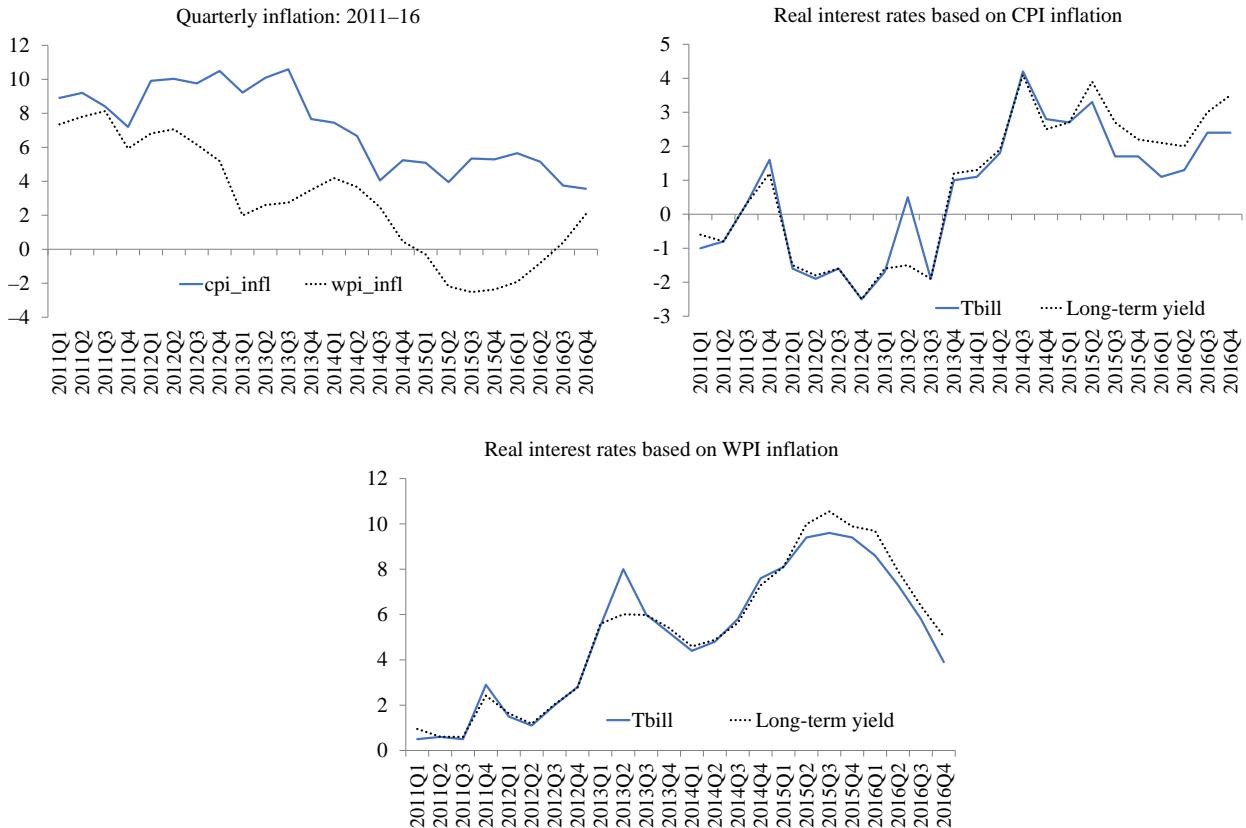


FIG. 2 Inflation and interest rates.

In this chapter, we estimate potential output by using the Hodrick–Prescott filter (HP filter). Many studies have compared the results of OG estimation by using different models in Indian context. [Bordlooi et al. \(2009\)](#) argue that unobserved components model yields the most efficient estimation for estimation of quarterly potential output. [Singh et al. \(2011\)](#) and [Kotia \(2016\)](#) also followed the same methodology for estimating potential output using Kalman filter. However, despite all the criticisms, HP filter is still widely accepted due to its simplicity. In this chapter, we estimate potential output by using the HP filter. Roughly speaking, HP filter decomposes a nonstationary time series, such as actual output, into Y_t^* a stationary cyclical component and Y_t^s a smooth trend component.

The observed correlations among the variables can be a good starting point to understand a *prima facie* relationship. [Fig. 3](#) provides comprehensive graphs based on correlation matrices of the variables in our model.

The correlation graphs shown in [Fig. 3](#) reveal some interesting facts. For example, OG does not show any correlation with retail inflation during our period of study. The simple correlation between private and public investment is very high and positive (even with public infrastructure and noninfrastructure) which signals that public investment may not be crowding out the private investment during this period. Nonfood credit, i.e., the credit flow to nonfood sectors also indicates a similar picture. The direction of causality, however, may be debated as some may argue that the decline in nonfood credit is independent of lack of demand for investment and not vice versa. Interest rates, whether short- or long-term highlight a significant positive relationship, suggesting that interest rates matter.

Finally, foreign portfolio capital flows show a low negative relationship with retail inflation dynamics and output gap. It does, however, show a low positive relationship with nonfood credit, public, and private investment. These *prima facie* findings regarding correlations among the variables provide further impetus to study more deeply, perhaps with sophisticated tools the dynamics of private investment, which is attempted in the next section.

4 Estimation and results

Our specification Eq. (1) incorporates both fiscal policy and monetary policy instruments relevant for encouraging private investment. We consider three versions of fiscal instruments (I_{pub}) as total public investment, public investment in infrastructure and noninfrastructure separately. This defines our three models. We also consider two versions of these three models with monetary policy variable real interest rates (i_r) based on short run and long run interest rates. Since public investment takes time to materialize, our models incorporate regressors for investments made two quarter before the current. Thus we have a comprehensive set up for testing the crowding controversy.

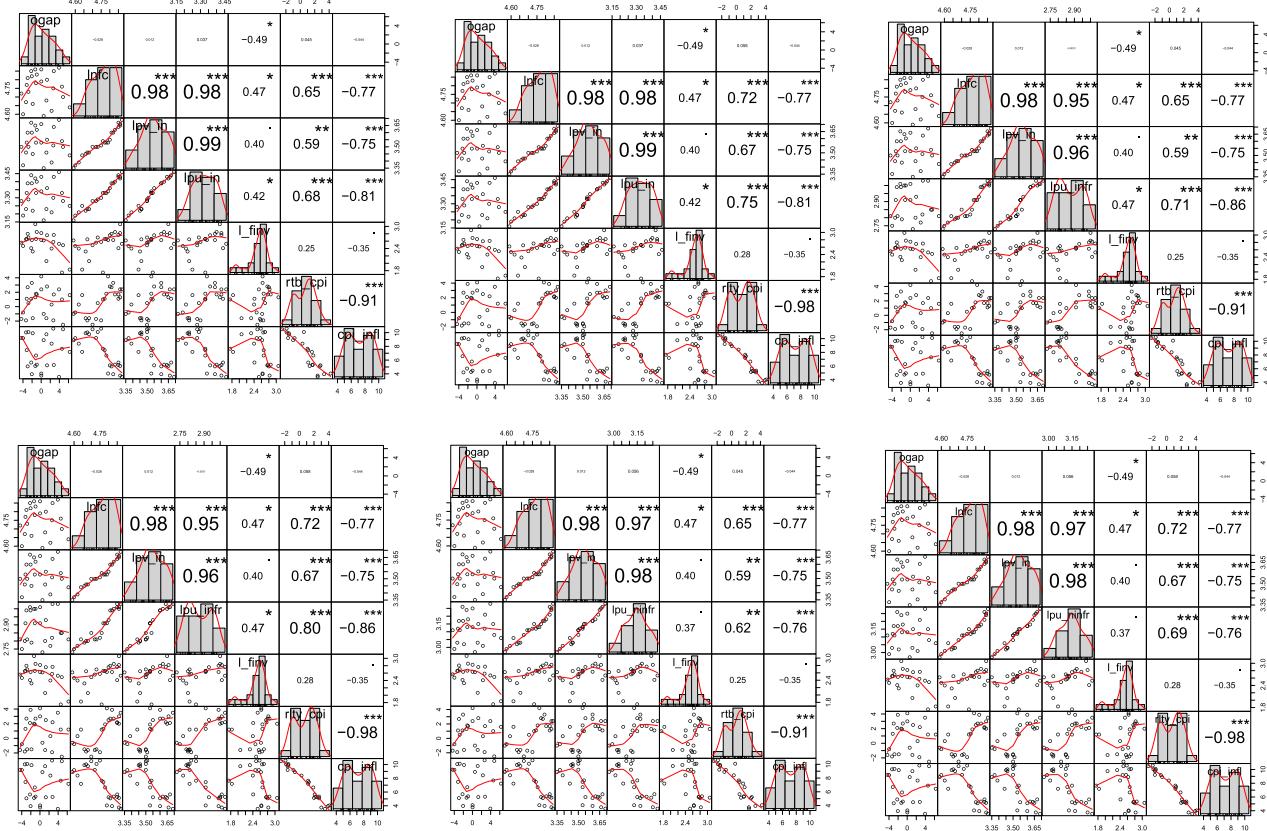


FIG. 3 Correlation matrices of the variables. Notes on graphs in Fig. 3: The distribution of each variable is shown on the diagonal. On the bottom of the diagonal: the bivariate scatter plots with a fitted line are displayed. On the top of the diagonal: the value of the correlation plus the significance level as stars. Each significance level is associated to a symbol: P -values ($0, 0.001, 0.01, 0.05, 0.1, 1$) have respective symbols (“****,” “***,” “**,” “*,” “.”). “Ogap” = output gap; “InfC” = log values of nonfood credit; “Ipv in” = log values of private investment; “Ipu in” = log values of public investment; “Ifinv” = log values of foreign investment; “Ipu_infr” = log values of public infrastructure investment; “Ipu_ninfr” = log values of public noninfrastructure investment; “rtb cpi” = cpi-based real treasury bills rate; “rly cpi” = cpi-based real long-term yield rate; “cpi infl” = cpi-based inflation.

Remark 1 The variables in Eq. (1) are a mixture of stationary and nonstationary variables, mostly as they are generated by government statistical agencies, without differencing or detrending. Since the underlying macroeconomic relation (among variables measured in levels) has theoretical support, it is obviously not spurious. Estimating it by ordinary least squares (OLS) is known to be “super-consistent” in the sense that its variance approaches zero at a fast rate T . Hence, even if some right-hand side variables in Eq. (1) are not exogenous, resulting “simultaneous equations bias” is of order \sqrt{t} , incapable of reversing super consistency of OLS.

Table 3 reports results for models when our i_r is the short-term interest rate. The column entitled “Model 1” regressors include current public investment and public investment made two quarters earlier. The column entitled “Model 2” regressors include similar two regressors referring to public infrastructure investment. The column entitled “Model 3” regressors include similar two regressors for public noninfrastructure investment. **Table 4** reports results for models when our i_r denotes long-term interest rate, but otherwise all column headings “Model j” are entirely analogous to those in **Table 3**.

We construct sophisticated confidence intervals for inference when the regression involves mixed variables using the meboot proposed by [Vinod \(2004\)](#) and has been developed extensively in [Vinod and López-de-Lacalle \(2009\)](#) and [Vinod \(2013\)](#). In the recent literature, researchers have discussed in detail the reliability of meboot methodology for time series inferences ([Chaiboonsri and Chaitip, 2013](#); [Lundholm, 2010](#); [Plasil, 2011](#); [Yalta, 2011](#); and among others). This chapter therefore, attempts to: (a) use the algorithm and tests to a short macroeconomic time series in the Indian context and (b) gain some insights regarding the developments in private investment in the economy in recent past. The description of various confidence intervals in **Tables 5–7** is as follows.^f

Simple percentile: The method is based on ordering $b_j^*, j=1, \dots, J$ values from the smallest to the largest as $b_j^*, j=1, \dots, J$. If $J=999$, $\alpha=0.05$, $(J+1)(\alpha/2)=25$ and $(J+1)(1-\alpha/2)=975$. Hence the “simple percentile” interval is given by the order statistics: $[b_{(25)}^*, b_{(975)}^*]$.

Boot percentile: This interval improves upon the “simple percentile” interval by working on a transformed scale to force the distribution of b^* to be symmetric, without knowing that transformation explicitly.

Norm: The “norm” interval uses a normal approximation to the distribution of “ b ” based on bootstrap estimates b^* of the bias and variance.

Basic: The “basic” confidence interval uses the following basic notion to better approximate the “norm” interval. Instead of directly using b^* to

^fFor more details, see [Vinod \(1983\)](#) and [Davison and Hinkley \(1997\)](#).

TABLE 3 Regression coefficient estimates using short-term interest rates

| | Dependent variable | | |
|---|---------------------------|----------------------|---------------------|
| | <i>Private investment</i> | | |
| | <i>Model 1</i> | <i>Model 2</i> | <i>Model 3</i> |
| Real T-bills rate | −0.008*** (0.001) | −0.008*** (0.002) | −0.004 (0.002) |
| Output gap | −0.002** (0.001) | −0.001 (0.001) | −0.001 (0.001) |
| Foreign investment | −0.022** (0.008) | −0.044*** (0.012) | −0.008 (0.017) |
| Nonfood credit | −0.271* (0.149) | 0.572*** (0.114) | 0.559** (0.247) |
| Public investment | 1.084*** (0.121) | | |
| Public investment, Lag2 | 0.394*** (0.100) | | |
| Public infrastructure investment | | 0.351*** (0.100) | |
| Public infrastructure investment, Lag2 | | 0.268** (0.093) | |
| Public Noninfrastructure investment | | | 0.575*** (0.184) |
| Public Noninfrastructure investment, Lag2 | | | 0.004 (0.167) |
| Constant | −0.016 (0.252) | −0.846** (0.303) | −0.910 (0.527) |
| Observations | 22 | 22 | 22 |
| R ² | 0.994 | 0.987 | 0.972 |
| Adjusted R ² | 0.992 | 0.981 | 0.961 |
| Akaike information criterion (AIC) | −145.435 | −126.616 | −110.590 |
| Bayesian information criterion (BIC) | −136.707 | −117.888 | −101.862 |
| Residual standard error (df=15) | 0.007 | 0.011 | 0.016 |

*P < 0.1; **P < 0.05; ***P < 0.01.

TABLE 4 Regression coefficient estimates using long-term interest rates

| | Dependent variable | | |
|---|---------------------------|----------------------|---------------------|
| | <i>Private investment</i> | | |
| | <i>Model 1</i> | <i>Model 2</i> | <i>Model 3</i> |
| Real long-term yield rate | −0.008*** (0.001) | −0.009*** (0.002) | −0.002 (0.003) |
| Output gap | −0.002** (0.001) | −0.001 (0.001) | −0.001 (0.002) |
| Foreign investment | −0.026** (0.009) | −0.051*** (0.012) | −0.008 (0.019) |
| Nonfood credit | −0.326* (0.184) | 0.537*** (0.112) | 0.506* (0.265) |
| Public investment | 1.134*** (0.152) | | |
| Public investment, Lag2 | 0.436*** (0.123) | | |
| Public infrastructure investment | | 0.451*** (0.108) | |
| Public infrastructure investment, Lag2 | | 0.265** (0.091) | |
| Public Noninfrastructure investment | | | 0.600*** (0.194) |
| Public Noninfrastructure investment, Lag2 | | | 0.013 (0.177) |
| Constant | −0.050 (0.312) | −0.947*** (0.303) | −0.763 (0.591) |
| Observations | 22 | 22 | 22 |
| R ² | 0.991 | 0.987 | 0.969 |
| Adjusted R ² | 0.988 | 0.982 | 0.957 |
| Akaike information criterion (AIC) | −136.388 | −127.497 | −108.123 |
| Bayesian information criterion (BIC) | −127.659 | −118.769 | −99.394 |
| Residual standard error (df=15) | 0.009 | 0.011 | 0.017 |

*P < 0.1; **P < 0.05; ***P < 0.01.

TABLE 5 Confidence intervals for “Model 1” defined in [Table 3](#)

| Variable | OLS | | Meboot | | | | | | | | HDR | |
|-------------------------------|--------|--------|-----------------------|--------|----------|--------|-----------|--------|------------|--------|--------|--------|
| | | | Simple percentile %Le | | Boot %Le | | Boot norm | | Boot basic | | | |
| | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% |
| Real treasury bills rate | -0.010 | -0.006 | -0.016 | 0.005 | -0.016 | 0.005 | -0.021 | 0 | -0.021 | 0 | -0.016 | 0.005 |
| Output gap | -0.003 | 0.000 | -0.003 | 0.002 | -0.003 | 0.002 | -0.004 | 0.001 | -0.004 | 0.001 | -0.003 | 0.002 |
| Foreign investment | -0.039 | -0.006 | -0.074 | 0.006 | -0.074 | 0.006 | -0.056 | 0.024 | -0.051 | 0.028 | -0.071 | 0.008 |
| Nonfood credit | -0.588 | 0.045 | -0.373 | 1.284 | -0.376 | 1.286 | -1.159 | 0.468 | -1.174 | 0.487 | -0.362 | 1.297 |
| Public investment | 0.826 | 1.342 | -0.309 | 1.22 | -0.313 | 1.23 | 0.725 | 2.188 | 0.674 | 2.216 | -0.311 | 1.201 |
| Public investment (Lag, 2) | 0.181 | 0.608 | -0.228 | 0.925 | -0.228 | 0.927 | -0.422 | 0.751 | -0.447 | 0.709 | -0.259 | 0.904 |

TABLE 6 Confidence intervals for “Model 2” defined in [Table 3](#)

| Variable | OLS | | Meboot | | | | | | | | | | HDR | |
|----------------------------------|--------|--------|-----------------------|--------|----------|--------|-----------|--------|------------|--------|--------|--------|-------|--------|
| | | | Simple percentile %Le | | Boot %Le | | Boot norm | | Boot basic | | | | | |
| | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% |
| Real treasury bills rate | -0.012 | -0.004 | -0.018 | 0.004 | -0.018 | 0.004 | -0.021 | 0.000 | -0.021 | 0.002 | -0.019 | -0.019 | | |
| Output gap | -0.003 | 0.002 | -0.003 | 0.002 | -0.003 | 0.002 | -0.003 | 0.002 | -0.003 | 0.002 | -0.003 | 0.002 | | |
| Foreign investment | -0.070 | -0.019 | -0.090 | -0.006 | -0.091 | -0.006 | -0.089 | -0.006 | -0.085 | 0.000 | -0.084 | -0.002 | | |
| Nonfood credit | 0.329 | 0.815 | -0.057 | 1.327 | -0.059 | 1.331 | -0.084 | 1.313 | -0.114 | 1.276 | -0.081 | 1.284 | | |
| Public infrastructure investment | 0.138 | 0.564 | -0.279 | 0.746 | -0.280 | 0.749 | -0.045 | 0.979 | -0.061 | 0.967 | -0.280 | 0.736 | | |
| Public infrastructure (Lag, 2) | 0.070 | 0.466 | -0.172 | 0.914 | -0.174 | 0.916 | -0.369 | 0.695 | -0.402 | 0.688 | -0.192 | 0.872 | | |

TABLE 7 Confidence intervals for “Model 1” defined in [Table 3](#)

| Variable | OLS | | Meboot | | | | | | | | HDR | |
|-------------------------------------|--------|--------|-----------------------|--------|----------|--------|-----------|--------|------------|--------|--------|--------|
| | | | Simple percentile %Le | | Boot %Le | | Boot norm | | Boot basic | | | |
| | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% | 2.50% | 97.50% |
| Real treasury bill rate | -0.009 | 0.001 | -0.014 | 0.006 | -0.014 | 0.006 | -0.015 | 0.005 | -0.015 | 0.005 | -0.013 | 0.006 |
| Output gap | -0.004 | 0.003 | -0.003 | 0.002 | -0.003 | 0.002 | -0.003 | 0.002 | -0.003 | 0.002 | -0.003 | 0.002 |
| Foreign investment | -0.046 | 0.029 | -0.062 | 0.027 | -0.062 | 0.027 | -0.050 | 0.039 | -0.049 | 0.040 | -0.061 | 0.028 |
| Nonfood credit | 0.032 | 1.086 | -0.223 | 1.315 | -0.225 | 1.318 | -0.058 | 1.506 | -0.067 | 1.476 | -0.245 | 1.285 |
| Public noninfrastructure investment | 0.183 | 0.967 | -0.177 | 1.097 | -0.177 | 1.103 | 0.018 | 1.284 | 0.000 | 1.280 | -0.177 | 1.087 |
| Public noninfrastructure (Lag, 2) | -0.352 | 0.360 | -0.348 | 0.737 | -0.349 | 0.738 | -0.747 | 0.304 | -0.772 | 0.314 | -0.345 | 0.738 |

approximate the unknown β , the observable deviations $b^* - b$ are likely to be better at approximating the unknown deviations $b - \beta$.

Tables 5–7 provide the confidence intervals under the “meboot” procedure for the three model specifications reported in Table 3. The basic descriptive statistics, graphs of ensembles, and “highest density regions” of the sampling distributions of coefficients of the variables (visually indicating the confidence intervals) are reported in the Appendix.

Our results show that “meboot” procedure can perform well in providing inference for short economic time series. Our results showed that it provides improvements in terms of narrower and balanced confidence interval bands in comparison to the standard OLS intervals. This is particularly useful for short time series where asymptotic inference is quite difficult to justify. The numerically constructed ensembles have the property of retaining the shape and autocorrelation and partial autocorrelation functions of the original time series data even for short time series. The highest density region (illustrated in the Appendix) of sampling distributions of estimated coefficients also indicate plausible confidence intervals.

Our confidence intervals continue to support “crowding in” of private investment through public investment for the period FY’ (fiscal year) 2011–FY’2016. Our findings are, thus, consistent with the recent literature using Indian data, which do not find crowding out effects of public investment on private investment. We find a significantly positive impact of credit cost reductions on corporate investment, albeit of a comparatively smaller magnitude than that of increases in public investment.

The direct crowding in effects of public infrastructure investment on corporate investment are evident in the lagged models. The estimated lagged coefficients (of two lags) reveal that 1% increase in infrastructure investment can lead to around 0.27 rise in private corporate investment in India. See Model 2 results in Tables 3 and 4. This signifies the spillover or second round effects of infrastructure investments in enhancing economic activity by boosting the confidence of private sector agents in their own new investment decisions. The instantaneous effect of public infrastructure investment on corporate investment is ranging between 35 and 45 basis points (bps) in all models. This implies a net reduction in project costs of private investment given the public infrastructure—an implicit net gain of an equivalent amount.

Output gap reflects macroeconomic uncertainties. As is evident from negative magnitude of the output gap variable, the macroeconomic uncertainties have negative impact on private corporate investment. The other finding of the study is that interest rates matter—both short term and long term. In other words, the cost of credit matters for corporate investment, though the magnitude of the impact is smaller than that of public investment variable. However, it does signify the sensitivity of interplay among infrastructure investment, interest rate changes, and private investment. Infrastructure investments usually require long gestation periods along with lower returns but higher social welfare gains.

If financed by state exchequer, they encourage interest rate sensitive private investment. Even if the government focuses on noninfrastructure investment, the significantly positive nonfood credit coefficient indicates that availability of credit instills a positive impact on private investment. During our time period, private investment may have galvanized to attain a larger share of resources, but did not essentially get crowded-out by the mere presence of public sector investment.

Our lagged models also reveal that the foreign investment (capital flows) variable was comparatively irrelevant in its effectiveness on corporate investment in India during the period of this analysis. The negative coefficient of foreign investment (which ideally could be argued to be positive to boost private investment in such equations of investment relations) also confirms to the concerns of the Economic Surveys highlighted earlier. Thus, amidst weak global demand and growth recovery, foreign investment flows could not boost or compliment private investment. Rather the uncertainty toward the stability in the flows of foreign capital had a negative bearing on the scale of private investment. The significant but opposite signs of nonfood credit indicate that mere quantity of credit may not be sufficient for enhancing private investment. The direct intervention of the state through focused infrastructure investment, thus, coupled with availability of credit can have stronger impact on private investment. This answers many riddles in Indian context—the role of the state is still a critical component for investment. Only by allocating resources for infrastructure, government encourages private investment.

The findings so far indicate that public investment matters and indeed leads to “crowding in” of private investment. Our initial correlation boxes indicated high and positive correlation between the two investments. A shortcoming here, as mentioned there, is that the direction of causality among the variables of interest is unknown from Pearson correlation coefficients. Accordingly, we use generalized correlation coefficients of [Vinod \(2014\)](#) which allow for non-linear relations among the variables evidently important from bivariate scatter diagrams of most pairs of variables in the model.

Since we also want to assess the causal directions, we use an exogeneity test statistic (or unanimity index) suggested by [Vinod \(2017\)](#) which allows determining the direction and strength of causal and exogenous variables. [Table 10](#) shows the results of sample causal path identifications. Many scientists have used this method including [Lister and Garcia \(2018\)](#) who call this “Vinod Causality.”

5 Data abbreviations and sources

The data source codes are: “Au” denotes authors’ calculations using public data. “CSM” denotes the Central Statistical Organization (CSO) and Ministry of Statistics and Programme Implementation. “RBI” denotes the Reserve Bank of India ([Table 8](#)).

TABLE 8 Data abbreviations used in computer implementations and the code for data sources

| No. | Code | Description | Source |
|-----|-------------|--|--------|
| 1 | nfc | Nonfood credit | |
| 2 | tbill | 91-Treasury bills rate | RBI |
| 3 | ltyield | 10 Years long-term yield | RBI |
| 4 | gdp_mp_curr | GDP mp at current prices | CSM |
| 5 | ogap | Output gap | Au |
| 6 | cpi_infl | Consumer price inflation | RBI |
| 7 | gfcf_cr | Gross fixed capital formation | CSM |
| 8 | pvt_in | Private investment | CSM |
| 9 | pub_in | Public investment | CSM |
| 10 | pub_infra | Public infrastructure investment | Au |
| 11 | pub_n_infra | Public noninfrastructure investment | Au |
| 12 | rtb | Real 91 treasury bills rate | Au |
| 13 | rlty | Real 10 years long-term yield rate | Au |
| 14 | lnfc | Log nonfood credit | RBI |
| 15 | lgfcf | Log gross fixed capital formation | RBI |
| 16 | lpv_in | Log private investment | RBI |
| 17 | lpu_in | Log public investment | RBI |
| 18 | lpu_infr | Log public infrastructure investment | RBI |
| 19 | lpu_ninfr | Log public noninfrastructure investment | RBI |
| 20 | l_pvt_in | Lag private investment | RBI |
| 21 | l_pv_in | Lag log private investment | RBI |
| 22 | finv | Foreign investment | RBI |
| 23 | l_finv | Log foreign investment | Au |
| 24 | rtb_cpi | Real short-term T-bill rate adjusted by CPI | Au |
| 25 | rlty_cpi | Long-term 10-year yield rate adjusted by CPI | Au |
| 26 | rtb_wpi | Real short-term T-bill rate adjusted by WPI | Au |
| 27 | rlty_wpi | Long-term 10-year yield rate adjusted by WPI | Au |

6 Causality results

Causal paths between 13 variables paired with private investment using symbols (up to seven characters in length) listed in [Table 9](#) are reported in [Table 10](#). They are obtained by using the R function “causeSummary” of the “generalCorr” package mentioned earlier.

The numbers in the column entitled “correlation” of [Table 10](#) are Pearson correlation coefficients. These are the usual symmetric correlation coefficients, measuring the nature of “linear” dependence between the variables named in the first two columns. Since all P -values are near zero except for output gap along line 3 of [Table 10](#) all relations in the table have statistically significantly nonzero Pearson correlation coefficients. However, the symmetry of the matrix of Pearson correlation coefficients means that they cannot suggest anything about the underlying causal directions.

When the value in the “strength” column of [Table 10](#) exceeds 15, the causal direction determination is strong enough to be believed as a

TABLE 9 Symbols using up to seven characters for variables in alphabetic order used in reporting causality paths in [Table 10](#)

| Symbol | Description |
|---------|--|
| FornInv | Log foreign investment |
| LongCPI | Long-term 10-year yield rate adjusted by CPI |
| LongWPI | Long-term 10-year yield rate adjusted by WPI |
| LongYld | 10-Year long-term yield |
| Ogap | Output gap |
| PbNnInf | Log public noninfrastructure investment |
| PubInfr | Log public infrastructure investment |
| PubInv | Log public investment |
| PvtInv | Log private investment |
| ReLngY | Real 10 years long-term yield rate |
| ReTbill | Real 91 treasury bills rate |
| ShrtCPI | Real short-term T-bill rate adjusted by CPI |
| ShrtRat | 91-Day treasury bills rate |
| ShrtWPI | Real short-term T-bill rate adjusted by WPI |

See [Table 8](#) for codes used in computer implementations and for data sources.

TABLE 10 Causal paths between selected variables using symbol in [Table 9](#)

| | Cause | Response | Strength | Corr. | P-value |
|----|---------|----------|----------|---------|---------|
| 1 | PvtInv | ShrtRat | 100 | -0.672 | 0.00032 |
| 2 | LongYld | PvtInv | 31.496 | -0.6862 | 0.00021 |
| 3 | Ogap | PvtInv | 100 | 0.0118 | 0.95626 |
| 4 | PvtInv | PublInv | 100 | 0.9868 | 0 |
| 5 | PvtInv | PublInfr | 100 | 0.9595 | 0 |
| 6 | PvtInv | PbNnInf | 100 | 0.9733 | 0 |
| 7 | PvtInv | ReTbill | 31.496 | 0.5854 | 0.00265 |
| 8 | PvtInv | ReLngY | 31.496 | 0.6745 | 3e-04 |
| 9 | PvtInv | FornInv | 31.496 | 0.4027 | 0.05103 |
| 10 | PvtInv | ShrtCPI | 31.496 | 0.5854 | 0.00265 |
| 11 | PvtInv | LongCPI | 31.496 | 0.6745 | 3e-04 |
| 12 | PvtInv | ShrtWPI | 100 | 0.7766 | 1e-05 |
| 13 | PvtInv | LongWPI | 100 | 0.8393 | 0 |

preliminary indicator of the true causal direction. Of course, the true direction is unknown in the absence of double blind-controlled experiments. This is the best we can assess using certain nonparametric kernel regressions and stochastic dominance of four orders from passively observed data.

It stands to reason that all variables except LongYld and Ogap along lines 2 and 3 of [Table 10](#) show that long-term yield and output gap influence the private investment, PvtInv, but all other variables are sensitive to independent variation in PvtInv DGP. It is interesting that “real” short-term or long-term interest rates adjusted by the consumer price inflation (CPI) or wholesale price inflation (WPI) give same causal path from PvtInv to various interest rates along rows 10–13.

Our results from meboot also indicate that public investment matters for private investment. The retail inflation-based real interest rate (CPI-based both short-term T-bill and long-term gov. yield) coefficients are negative and significant when aggregate public investment is included as an explanatory variable.

The rationale of using both price and quantity variables in the equation is to test for McKinnon hypothesis: whether cost or quantity of credit matter for private corporate investment. Although the two are inversely related in administrative regimes, our data cover mostly deregulated regime in India. Regarding the real interest variable, debate exists whether to use WPI or CPI deflator. Our results focus on the latter for brevity, since our causality path analysis indicates that both behave in a similar fashion. A limitation here is that we did not use the investment deflator from the national accounts data.

Another limitation is that we used ex post interest rate (based on backward-looking inflation) for our estimation, rather than an ex ante rate (based on forward-looking inflation expectations). Since investment is in its very essence a forward-looking activity, an ex ante rate would seem more appropriate. However, a long-enough time series for “long-term inflation expectations” remains unavailable for the new macro series regime in India, even if we were to use consensus inflation forecasts or those by the RBI professionals.

[Chakraborty \(2016\)](#) shows that the link between public investment and private corporate investment is positive and significant in models with and without lags. Hence models presented in our chapter incorporate both coincident and two-quarter lag effects on private corporate investment.

7 Conclusion

The literature aimed at finding determinants of private investment relies on two types of models: autoregressive distributed lag (ARDL) and vector autoregression (VAR). Both these models are quite useful for time series estimations in case the focus is on long-term relations. These models try to transform data variables to stationary variables by differencing or detrending.

In this chapter, we consider an analysis of causal paths to assess the endogeneity of various stationary and nonstationary macroeconomic variables measured in their levels, without differencing or detrending to convert them into stationary variables. Our causal path analysis using the R package “generalCorr” shows that private investment as a data generating process has independent variation which drives the variation in public infrastructure and noninfrastructure investments and also the variation in long-term government bond rates. It highlights the importance of private investment as a driving force for the growth of Indian economy and difficulties in choosing policies to influence it.

The estimated directions of causal paths make some variables endogenous. However, [Remark 1](#) following our specification of Eq. (1) above

shows that despite endogeneity, our model can be reliably estimated by OLS, which is known to be superconsistent. Hence we need not be concerned with the “simultaneous equations bias” induced by potential endogeneity.

The inference for mixed model specifications such as ours is possible because of meboot for constructing confidence intervals based on short macroeconomic time series and explore regression models for determination of private investment in India. The methodology allows overcoming the unit root and structural change pretests while ruling out the need for any (detrending or differencing) transformations of original time series, merely for statistical nicety of ensuring the stationarity assumption.

Our results for fiscal years 2011–16 indicate evidence in support of “crowding in” of private investment through public investment. We find that public infrastructure investment is significant in determining private investment and that a low interest rate encourages private corporate investment. When the government commits resources for infrastructure, it creates incentives for enhancing private investment.

Private corporate investment is often cyclical, whereby investment booms are followed by recessions, reflecting among other issues the fact that firm level capacity utilization or capacity addition are often bulky, expensive and uncertain. Our time period covers mostly a recessionary phase of the investment cycle following a modest expansion. In the absence of data to cover many business cycles, we capture some aspects of cyclical behavior by including the “output gap” variable in the model. The public policy implication of our chapter is that the government should remove the infrastructure and bureaucratic bottlenecks in the economy by enhancing “ease of doing business” in India. The recent initiative by the government, for instance, “Make in India,” is also a promising policy initiative to promote greater private investment.

Many macroeconomic empirical studies estimate equations similar to (1) involving short time series data in levels. These researchers can treat our R code at^g as a template for the use of R packages “meboot” (for time series bootstrap statistical inference) and “generalCorr” (for estimation of causal paths).

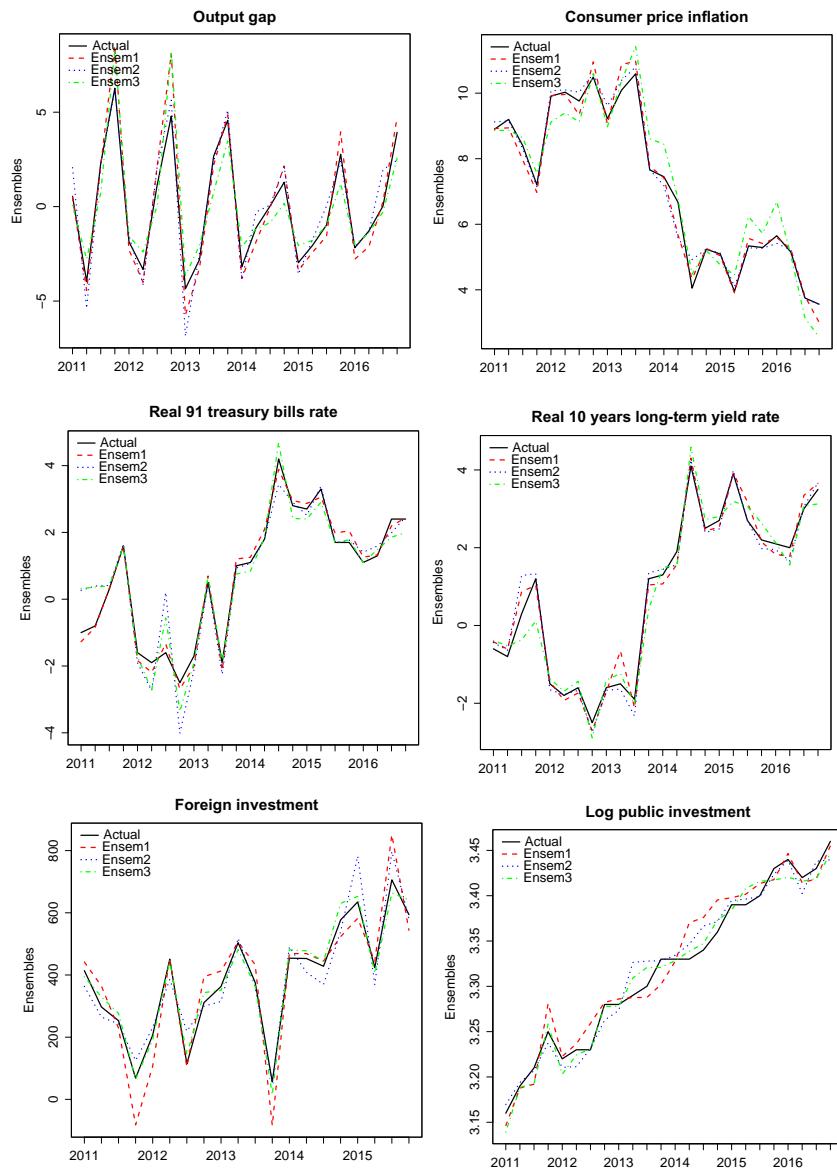
Appendix

The basic descriptive statistics (Table A.1) and Figs. A.1 and A.2 related to bootstrap confidence intervals are provided here, as mentioned in Section 4.

^gSee <http://www.fordham.edu/economics/vinod/vkcR.zip>.

TABLE A.1 Descriptive statistics

| | Private inv. (INR bn) | Public inv. (INR bn) | Public infra. inv. (INR bn) | Pub. noninfra. inv. (INR bn) | Output gap | Nominal treas. bills % | Nominal G-sec yield (%) | Nonfood credit (INR bn) | Foreign inv. (INR bn) |
|-----------|--------------------------|-------------------------|--------------------------------|---------------------------------|------------|------------------------|-------------------------|----------------------------|--------------------------|
| Min. | 22.32 | 14.59 | 5.32 | 9.27 | -4.22 | 7.03 | 7.52 | 389.90 | 56.00 |
| First Qu. | 28.43 | 17.09 | 6.30 | 10.78 | -2.31 | 7.85 | 7.92 | 462.30 | 286.00 |
| Median | 32.32 | 19.62 | 6.81 | 12.88 | -0.95 | 8.21 | 8.27 | 548.10 | 419.00 |
| Mean | 31.74 | 19.89 | 7.09 | 12.79 | -0.12 | 8.21 | 8.24 | 543.80 | 384.20 |
| Third Qu. | 35.05 | 21.29 | 7.71 | 13.76 | 2.13 | 8.60 | 8.56 | 620.90 | 466.50 |
| Max. | 39.75 | 26.50 | 9.40 | 17.11 | 6.93 | 10.56 | 8.90 | 704.50 | 706.00 |
| SD | 4.84 | 3.36 | 1.15 | 2.25 | 3.29 | 0.56 | 0.73 | 97.14 | 181.11 |

**FIG. A.1** Maximum entropy ensembles of selected variables.

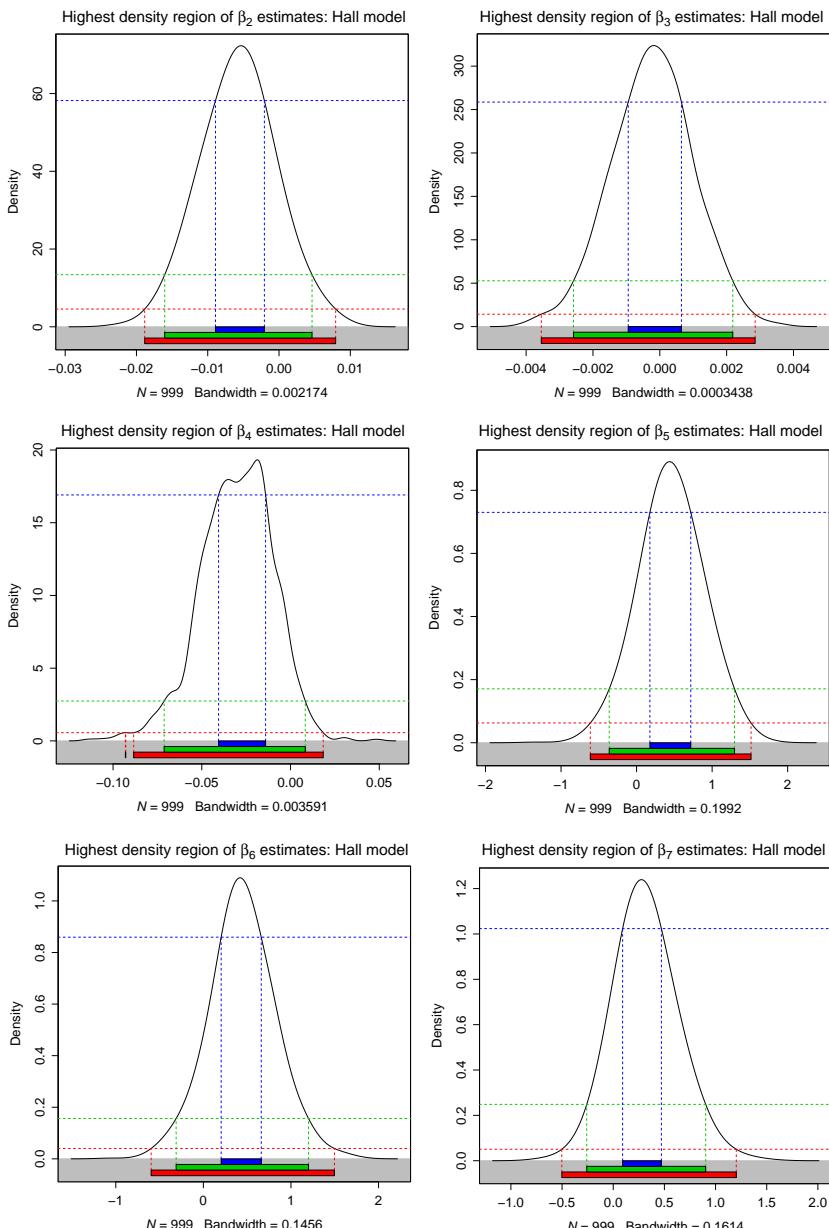


FIG. A.2 High density region (HDR) for coefficients—model 1.

References

- Alesina, A., Ardagna, S., Perotti, R., Schiantarelli, F., 2002. Fiscal policy, profits and investment. *Am. Econ. Rev.* 92 (3), 571–589.
- Aschauer, D.A., 1989. Does public capital crowd out private capital? *J. Monet. Econ.* 24 (2), 171–188.
- Bahal, G., Raissi, M., Tulini, V., 2015. Crowding-out or crowding-in? Public and private investment in India. IMF Working Paper 264.
- Blejer, M.I., Khan, M.S., 1984. Government policy and private investment in developing countries. *IMF Staff. Pap.* 31 (2), 379–403.
- Bordlo, S., Das, A., Jangili, R., 2009. Estimation of potential output in India. *RBI Occas. Pap.* 30 (2), 37–73.
- Buiter, W., 1977. Crowding out and the effectiveness of fiscal policy. *J. Public Econ.* 7 (3), 309–328.
- Cebula, R.J., 1978. An empirical analysis of the ‘crowding out’ effect of fiscal policy in the United States and Canada. *Kyklos* 31 (3), 3424–3436.
- Chaiboonsri, C., Chaitip, P., 2013. A boundary analysis of ICT firms on thai-land stock market: a maximum entropy bootstrap approach and highest density regions (HDR) approach. *Int. J. Comput. Econ. Econ.* 3 (1/2), 14–26.
- Chakraborty, L., 2007. Fiscal deficit, capital formation, and crowding out in India: evidence from an asymmetric VAR model. *Economics Working Paper Archive. The Levy Economics Institute*, New York, WP 518.
- Chakraborty, L., 2012. Interest rate determination in India: empirical evidence on fiscal deficit-interest rate linkages and financial crowding out. *Economics Working Paper Archive, The Levy Economics Institute*, New York, WP 744.
- Chakraborty, L., 2016. *Fiscal Consolidation, Budget Deficits and the Macroeconomy*. Sage Publications, UK, ISBN: 9789351509899.
- Chhibber, A., Kalloor, A., 2016. Reviving private investment in India: determinants and policy levers. *NIPFP Working Paper Series. National Institute of Public Finance and Policy*, New Delhi. WP 181.
- Dash, P., 2016. The impact of public investment on private investment: evidence from India. *J. Decis. Mak.* 41 (4), 288–307.
- Davison, A.C., Hinkley, D.V., 1997. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.
- Efron, B., 1979. Bootstrap methods: another look at jackknife. *Ann. Stat.* 7, 1–26.
- Erenburg, S.J., 1993. The real effects of public investment on private investment. *Appl. Econ.* 25 (6), 831–837.
- Greene, J., Villanueva, D., 1990. Private investment in developing countries: an empirical analysis. *IMF Staff. Pap.* 38 (1), 33–58.
- Kotia, A., 2016. An Unobserved Components Phillips Curve in an Emerging Market Economy: The Case of India. Available at SSRN: <https://ssrn.com/abstract=2765176>.
- Lister, B.C., Garcia, A., 2018. Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proc. Natl. Acad. Sci. U. S. A.* 15, 1–10. <http://www.pnas.org/content/early/2018/10/09/1722477115.full.pdf>.
- Lundholm, M., 2010. Are inflation forecasts from major swedish forecasters biased? *Working Paper, Department of Economics, Stockholm University*.
- Mallick, J., 2016. Effects of government investment shocks on private investment and income in India. *Indian Council for Research on International Economic Relations*, New Delhi, WP 315.

- Nelson, C.R., Plosser, C.I., 1982. Trends and random walks in macroeconomic time series. *J. Monet. Econ.* 10, 139–162.
- Ostrosky, A., 1979. An empirical analysis of the crowding out effect of fiscal policy in the United States and Canada: comments and extensions. *Kyklos* 32 (3), 497–522.
- Parker, K., 1995. The behaviour of private investment. IMF Occasional Paper 134, International Monetary Fund, Washington, DC.
- Plasil, M., 2011. Potential product, output gap and uncertainty rate associated with their determination while using the hodrick-prescott filter. *Polit. Ekon.* 59, 490–507.
- Pradhan, B.K., Ratha, D.K., Sarma, A., 1990. Complementarity between public and private investment in India. *J. Dev. Econ.* 33 (1), 101–116.
- Ramirez, M.D., 1994. Public and private investment in Mexico, 1950–90: an empirical analysis. *South. Econ. J.* 61 (1), 1–17.
- Shafik, N., 1992. Modeling private investment in Egypt. *J. Dev. Econ.* 39, 263–277.
- Singh, B., Kanakaraj, A., Sridevi, T., 2011. Revisiting the empirical existence of the Phillips curve for India. *J. Asian Econ.* 22, 247–258.
- Sundararajan, V., Thakur, S., 1980. Public investment, crowding out and growth: a dynamic model applied to India and Korea. *IMF Staff Pap.* 27 (4), 814–855.
- Tanzi, V., 1985. Fiscal deficits and interest rates in the United States: an empirical analysis. *IMF Staff Pap.* 32 (4), 551–561.
- Toda, H.Y., Yamamoto, T., 1995. Statistical inference in vector autoregressions with possibly integrated processes. *J. Econ.* 66 (1–2), 225–250.
- Tun, W.U., Wong, C., 1982. Determinants of private investment in developing countries. *J. Dev. Stud.* 19 (1), 19–36.
- Vinod, H.D., 1983. Stable and low public utility rates by wiener-hopf optimization. In: 1982 Proceedings of the Business and Economics Section. American Statistical Association, Washington, DC, pp. 369–374.
- Vinod, H.D., 2004. Ranking mutual funds using unconventional utility theory and stochastic dominance. *J. Empir. Financ.* 11 (3), 353–377.
- Vinod, H.D., 2006. Maximum entropy ensembles for time series inference in economics. *J. Asian Econ.* 17 (6), 955–978.
- Vinod, H.D., 2009. Stress testing of econometric results using archived code for replication. *J. Econ. Soc. Meas.* 34 (2–3), 205–217.
- Vinod, H.D., 2013. Maximum entropy bootstrap algorithm enhancements. SSRN eLibrary. <http://ssrn.com/paper=2285041>.
- Vinod, H.D., 2014. Matrix algebra topics in statistics and economics using R. In: Rao, M.B., Rao, C.R. (Eds.), *Handbook of Statistics: Computational Statistics With R*, vol. 34. Elsevier Science, North Holland, New York, pp. 143–176. Ch. 4.
- Vinod, H.D., 2017. Causal paths and exogeneity tests in Generalcorr package for air pollution and monetary policy. SSRN eLibrary. <http://ssrn.com/paper=2982128>.
- Vinod, H. D., Chakraborty, L. S., Karun, H., 2014. If deficits are not the culprit, what determines Indian interest rates? An evaluation using the maximum entropy bootstrap method. Working paper no. 811, Levi Economics Institute of Bard College. URL http://www.levyinstitute.org/files/download.php?file=wp_811.pdf&pubid=2104
- Vinod, H.D., López-de-Lacalle, J., 2009. Maximum entropy bootstrap for time series: the meboot R package. *J. Stat. Softw.* 29 (5), 1–19. <http://www.jstatsoft.org/v29/i05/>.
- Yalta, A.T., 2011. Analyzing energy consumption and gdp nexus using maximum entropy bootstrap: the case of Turkey. *Energy Econ.* 33 (2011), 453–460.

Chapter 6

High-mixed frequency forecasting methods in R—With applications to Philippine GDP and inflation

Roberto S. Mariano* and Suleyman Ozmucur

Department of Economics, University of Pennsylvania, Philadelphia, PA, United States

*Corresponding author: e-mail: mariano@upenn.edu

Abstract

Recognizing the need to utilize high-frequency indicators for more up-to-date forecasts, this chapter surveys alternative modeling approaches to combining mixed frequency data for forecasting purposes. The models covered in this chapter include data-parsimonious (but more computer-demanding) models such as the mixed-frequency dynamic latent factor model (MF-DLFM) as well as more data-intensive ones like the current quarterly model (CQM) and mixed data sampling (MIDAS) regressions. In all these models, the fact that the data set is of mixed frequencies raises technical issues in the estimation and forecasting phases of the exercise. In the case of MF-DLFM, the additional feature of unobserved common factors introduces additional complications in implementing the estimation and simulation strategy based on the derived observable state-space formulation of the model.

The alternative models are estimated and constructed using Philippine data, to forecast GDP growth and inflation in the Philippines. For this numerical exercise, 10 monthly indicators are used for quarterly real GDP and 9 monthly indicators for the quarterly GDP deflator. The whole empirical analysis is implemented in R—starting from using R to access Philippine data from Philippine and international data sources to analyze the statistical properties of Philippine real GDP and GDP deflator and culminating in the estimation of the alternative forecasting models, where numerous variations of MIDAS are explored.

As the next step in this research, it would be particularly important to compare the forecasting performance of the alternative procedures that are surveyed in this chapter. A more comprehensive study of this type will be presented in a future sequel to this chapter. Indicative comparison results that have just been recently reported point to the potentially superior performance of MF-DLFM for forecasting GDP growth, while

for forecasting inflation, the performance of MF-DLFM is not significantly better than MIDAS. More work is required for a more definitive conclusion on this issue—requiring further analysis and empirical applications, especially in expanding the performance analysis to cover the wider span of alternative forecasting models and variations of MF-DLFM and MIDAS surveyed in this chapter. Dynamic simulations for multiperiod forecasting also should be considered, as well as more refinements in the estimated models, especially the dynamic latent factor models, and extension of the analysis to other countries, especially in Southeast Asia.

Keywords: Forecasting in R with mixed-frequency data, Nowcasting, Dynamic latent factor models, MIDAS, Bridge equations, Current-quarter modeling, Kalman filter, Forecasting Philippine GDP and inflation, Principal components, ARIMA, VAR

1 Introduction

Combining mixed high-frequency data—e.g., quarterly, monthly, weekly, even daily for short-term forecasting has generated considerable renewed interest. The timely and statistically efficient use of “breaking news” is critical in a wide range of disciplines, where harnessing high-frequency indicators for more up-to-date forecasts and assessment is particularly important. Two disciplines of note are financial econometrics and macroeconomic forecasting, as data information in these fields have become richer, more diversified, nonstandard, and available at different and higher frequencies. This is especially so for government policy planners who need to monitor the state of the economy in real time, as well as financial managers and analysts.

This chapter investigates the technical and practical issues involved in the use of data at mixed frequencies (quarterly and monthly and, possibly, weekly and daily) to forecast monthly and quarterly economic activity in a country. The analysis considers alternative high-frequency forecasting models for GDP growth and inflation, utilizing indicators that are observable at different frequencies. The chapter focuses in particular on dynamic time-series models that involve latent factors and compares the forecasting performance of this approach with more commonly used data-intensive methods that have been developed in applications in the United States and Europe—in particular, Mixed Data Sampling (MIDAS) Regression and Current Quarter Modeling (CQM) with Bridge Equations. While these alternatives are mostly data-intensive, the dynamic latent factor modeling with mixed frequencies presents a parsimonious approach which depends on a much smaller data set that needs to be updated regularly. But it also faces additional complications in methodology and calculations as mixed-frequency data are included in the analysis.

The alternative models are estimated and constructed using Philippine data, to forecast GDP growth and inflation in the Philippines. For this numerical exercise, 10 monthly indicators are used for quarterly real GDP and

9 monthly indicators for the quarterly GDP deflator. The whole empirical analysis is implemented in R (see, [Bennett and Hugen, 2016](#); [Heiss, 2016](#); [Hyndman and Athanasopoulos, 2018](#); [R Development Core Team, 2018](#); [Shumway and Stoffer, 2017](#); [Vinod, 2011](#))—starting from using R to access Philippine data from Philippine and international data sources to analyzing the statistical properties of Philippine real GDP and GDP deflator and culminating in the estimation of the alternative forecasting models, where numerous variations of MIDAS are explored.

As the next step in this research, it would be particularly important to compare the forecasting performance of the alternative procedures that are surveyed in this chapter. A more comprehensive study of this type will be presented in a future sequel to this chapter. Indicative comparison results that have just been recently reported point to the potentially superior performance of MF-DLFM for forecasting GDP growth, while for forecasting inflation, the performance of MF-DLFM is not significantly better than MIDAS. More work is required for a more definitive conclusion on this issue—requiring further analysis and empirical applications, especially in expanding the performance analysis to cover the wider span of alternative forecasting models and variations of MF-DLFM and MIDAS surveyed in this chapter. Dynamic simulations for multiperiod forecasting also should be considered, as well as more refinements in the estimated models, especially the dynamic latent factor models, and extension of the analysis to other countries, especially in Southeast Asia.

[Section 2](#) summarizes the alternative forecasting models studied in this chapter. [Section 3](#) goes into the application to forecasting real GDP and inflation in the Philippines and computer implementation in R. Specific details about the estimated forecasting models and their implementation in R are provided in [Section 4](#). Comparison of forecasts and concluding remarks are in [Section 5](#).

2 Alternative forecasting models in this study

In general, the data analyst may encounter situations with a mixed-frequency data set, which may include quarterly, monthly, weekly, and daily observations. In this chapter, the target variables are the year-on-year growth rates of real GDP and the GDP deflator; these are available quarterly. On the other hand, all the indicator variables are available monthly. Note that the forecasting procedures implemented here can be adapted to more general situations where the indicator variables come in mixed frequencies.

The alternative forecasting models we consider here can be labeled as “quarterly” or “monthly,” according to the basic or underlying frequency that is explicitly modeled. For the quarterly models, observed quarterly values of the target variables are directly utilized, while observations for the monthly indicators are aggregated over the quarter. For example, for stock variables,

averages are calculated over the quarter, sums are utilized for flow variables, and growth rates are calculated from the aggregated series.

A monthly model, on the other hand, treats all the data series (target or indicator) as generated at the highest frequency (monthly, in our case), but some of the data points are not observed. Variables observed at the low frequency (quarterly) are treated as having periodically missing or unobserved data points, available only at the end month of the quarter. Estimation procedures are then implemented to take account of the presence of systematically missing observations. Note that an estimated monthly model also would provide forecasts of the target variables disaggregated at the high frequency (monthly, in our Philippine example).

2.1 Quarterly models

The following quarterly models are covered in this study:

1. Benchmark (Vector) Autoregressive Moving Average Processes (no indicators used)

$$Y_{tq} \sim \text{ARMA}(p,r) \text{ or VARMA}(p,r)$$

2. Bridge Equations (expanding the benchmark by introducing indicator variables, possibly w/lags, as additional explanatory variables)

$$Y_{tq} \sim [\text{ARMA}(p,r), Z_{tq}]$$

3. Bridge—PCA (principal components)

$$Y_{tq} \sim [\text{ARMA}(p,r), \text{PC}(Z_{tq})]$$

4. Current Quarterly Model (CQM)—bridge modeling for high-frequency updates of forecasts of GDP and its components. Here the objective is timely forecast of GDP and its components in the national income accounts, typically available quarterly. “Bridge” equations are used, relating GDP components to observable quarterly and monthly “indicator” variables. Monthly observations are averaged over the quarter, with updates as more monthly observations become available. To forecast the monthly and quarterly indicators, ARIMA models are used. If no indicators are available, an ARIMA model would be estimated for the GDP component itself.

CQM with bridge equations for the United States was researched extensively by Lawrence Klein—e.g., in [Klein and Sojo \(1987, 1989\)](#), [Klein and Park \(1993, 1995\)](#), [Klein and Ozmucur \(2001, 2002, 2004, 2008\)](#), [Mariano and Tse \(2008\)](#), and [Mariano and Ozmucur \(2018\)](#) in [Pauly \(2018\)](#). Now, CQM models have been developed for updating quarterly forecasts in various countries like Turkey ([Ozmucur, 2009](#)), Japan ([Inada, 2005](#)), Mexico ([Coutino, 2005](#)), Russia ([Klein et al., 2003, 2005](#)), and China ([Klein and Mak, 2005](#)).

2.2 Monthly models

The following monthly models are covered in this study

1. Monthly VAR using averages or cubic splines to “fill in the blanks”—namely, estimate missing monthly observations
2. Mixed-Frequency Vector Autoregressive (MF-VAR)

$$Y_{tm} \sim [VAR(p), Z_{tm}]$$

This has a state-space model formulation and Kalman filtering methods can be used to estimate the model and calculate forecasts at the highest frequency—e.g., see [Harvey \(1989\)](#).

3. Mixed Data Sampling (MIDAS) Regressions—alternative lag structures and alternative variations, including combinations with dynamic factor models
4. Mixed frequency dynamic latent factor model (MF-DLFM)

2.3 MIDAS regressions

Typical bridge equation modeling relates a quarterly variable to 3-month averages of monthly variables. This implicitly imposes a restriction on coefficients for the months of the quarter and consequently introduces asymptotic biases and inefficiencies—[Ghysels, 2013](#).

In contrast, MIDAS estimates a monthly regression of GDP on monthly (and possibly quarterly) indicators using parsimonious distributed lags to represent missing observations. The Initial reference is [Ghysels et al. \(2004\)](#), with early applications in finance, now also used to forecast macroeconomic time series. Since its introduction, this modeling approach has been used extensively in the mixed-frequency forecasting literature and has been enhanced with numerous variations—as described, for example, in [Ghysels \(2016a,b\); Ghysels et al. \(2007\)](#); and [Ghysels and Marcellino \(2018\)](#).

For implementation, MIDAS applies a more parsimonious parametrization of distributed lag structures to model the relation of GDP to current and lagged indicators at the monthly frequency, so that the basic model can be expressed as.

$$Y_{tm} \sim DL(Z_{tm}) + \text{error}$$

The estimation method is nonlinear least squares using actual observed data at mixed frequencies. Early examples of lag structures used in MIDAS include

1. Unrestricted (but truncated)
2. Step Function (equal weights for months of same quarter, truncated)
3. Polynomial Almon Lag
4. Exponential Almon

$$c_k = \exp(\theta_1 k + \theta_2 k^2) / \sum_k \exp(\theta_1 k + \theta_2 k^2)$$

5. Beta Lag

$$c_k = f(k/K; a, b) / \sum_k f(k/K; a, b)$$

$$f(x; a, b) = x^{a-1} (1-x)^a G(a+b) / [G(a)G(b)]$$

The following are some extensions and variations of MIDAS which we experiment with in our Philippine example:

1. Autoregressive MIDAS—ADL-MIDAS (add lags of the dependent variable as additional regressors)
2. MIDAS—MF-DLFM or Factor MIDAS (include latent factors in the equation).

2.4 Mixed-frequency dynamic latent factor models (MF-DLFM)

The underlying philosophy is that macroeconomic fluctuations are driven by a small number of common shocks or factors and an idiosyncratic component peculiar to each economic time series. The seminal papers are [Sargent and Sims \(1977\)](#) and [Stock and Watson \(1989\)](#). Earlier works (e.g., Stock and Watson) develop single factor models to construct composite indices of economic activity based on a handful of coincident indicators. More recent studies use the model to extract unobserved common factors from a large collection of observable indicator variables. More recently, the approach was revived for forecasting purposes in the United States and larger European countries—[Foroni and Marcellino \(2012, 2013\)](#).

Another (related) application has dealt with combining mixed frequencies in constructing composite indices—e.g., [Mariano and Murasawa \(2003\)](#), [Aruoba et al. \(2009\)](#).

The estimated factor model, properly validated, also may be used to forecast macroeconomic variables of interest at the highest frequency, e.g., [Liu and Hall \(2001\)](#), [Mariano and Murasawa \(2010\)](#).

In brief, the underlying model consists of two parts. The first explains the dynamics of the target and indicator variables depending on own lags, unobservable common factor(s), and possibly, observable exogenous variables. The second part explains the behavior of the latent common factor(s) in terms of their own joint dynamics and possibly, interactions with observable indicators. The system may also have other observable exogenous variables that serve as indicators for the latent common factors.

A similar modeling approach is used in [Mariano and Murasawa \(2003, 2010\)](#) in constructing an improved coincident economic index indicator in the United States using mixed frequencies (quarterly and monthly), as well as in [Aruoba et al. \(2009\)](#) in constructing a “real-time” (daily) business conditions index for the United States, using four indicators (quarterly, monthly, weekly, daily).

To render the analysis implementable, we have to cope with the two confounded complications of missing data observations as well as unobserved common factors. One solution is to derive from the underlying model, a

state-space formulation with measurement and state equations involving only fully observed variables, latent state variables, predetermined variables, and measurement and transition shocks.

“Missing” observations need to be factored in constructing the observation matrices in the state-space formulation and one needs to distinguish treatment of stock and flow variables. Also, the linear state-space formulation is only an approximation to the true relationship—nonlinear filtering procedures, typically through stochastic simulations, would be needed to get an exact solution; but linear approximations may suffice.

Details on the specific expressions for the variables and parameters in the measurement and state equations depend on the mixed frequencies present in the model. And they get more complicated and more computer intensive as higher and higher frequencies get involved.

Kalman filtering procedures can be applied to reestimate unknown parameters in this state-space formulation and perform signal extraction to calculate estimates of the latent factor. This Kalman filtering approach needs to be adapted to special complicating features of the problem. In particular, using mixed-frequency data for the indicators introduces missing data in the “measured” variables. Also, additional attention is needed and further complications in calculations arise when dealing with indicators that are flow variables.

Details for formulating the “observable” state space model are in [Harvey \(1989\)](#), [Mariano and Murasawa \(2003, 2010\)](#), and [Aruoba et al. \(2009\)](#).

3 Application to forecasting Philippine GDP and inflation and computer implementation in R

3.1 Getting data with R

In addition to quarterly real GDP (GDP), and quarterly GDP deflator (DEF), there are 17 higher frequency indicators. Two variables are available daily (stock prices, and exchange rates). Their monthly averages of daily figures are used in this chapter. For a forecasting exercise, daily averages may definitely be utilized. For example, on October 29th, monthly figure for October is not available, but average of first 29 days may be used before the month for October figure is released. One may also utilize a MIDAS model to find the relationship with a monthly (or quarterly) and a daily series.

There are 10 monthly indicators for real GDP: industrial production index (y-o-y growth) (X01), merchandise imports (y-o-y growth) (X02), merchandise exports (y-o-y growth) (X03), real government expenditure (y-o-y growth) (X04), real money supply (M1) (y-o-y growth) (X05), world trade volume (y-o-y growth) (X06), real stock price index (y-o-y growth) (X07), real exchange rate (y-o-y growth) (X08), time deposit rate—savings deposit rate (X09), treasury bill rate (91-day)—US treasury bill rate (3-month) (X10).

There are 9 monthly indicators for quarterly GDP deflator: consumer price index (y-o-y growth) (X11), producer price index (y-o-y growth) (X12),

wholesale price index, Metro Manila (y-o-y growth) (X13), retail price index (y-o-y growth) (X14), exchange rate (y-o-y growth) (X15), money supply (M1) (y-o-y growth) (X16), world Inflation (based on CPI) (y-o-y growth) (X17), time deposit rate—savings deposit rate (X09), treasury bill rate (91-day)—US treasury bill rate (3-month) (X10).

Real GDP and nominal GDP figures are calculated and released by the Philippine Statistics Authority (<<http://www.psa.gov.ph>>). These are in current and constant 2000 prices. The latest release was on August 9, 2018 covering the period 1998Q1 to 2018Q2 (<http://www.psa.gov.ph/nap-press-release/data-series>). They also provide quarterly data for the 1981Q1-1997Q4, and annual data for the 1946–2010 period to link and extend the series. However, these are in table formats with empty columns. It is not easy to construct a time-series data from such tables.

It is possible to get data from FRED (Federal Reserve Bank of St. Louis's data bank with over a half a million series). FRED seems to have about 300 series for the Philippines, but close examination reveals that quite a few quarterly series are discontinued (<https://fred.stlouisfed.org/categories/32788>). There are about 300 low frequency (annual or 5-year) series from the World Bank. Residential property prices by the Bank of International Settlements are available quarterly from the first quarter of 2008. There are 4 monthly series. Two of these series are by the Bank of International Settlements and on effective exchange rate index and covering the post January 1994 period. The other two, United States exports and imports to Philippines, are from the US Bureau of Economic Analysis. It will be very useful if more variables are added to FRED because the R package “pdfetch” is a very convenient way to import data from FRED and the World Bank among others (Reinhart, 2017).

```
> phbis<- pdfetch_FRED(c("NBPHBIS", "RBPHBIS"))
> rbind(c(head(phbis,3),tail(phbis,3)))
   NBPHBIS RBPHBIS
1994-01-31  163.46  98.40
1994-02-28  162.39  98.01
1994-03-31  162.33  97.36
2018-07-31   93.91 102.93
2018-08-31   94.68 104.30
2018-09-30   93.62 103.75
```

3.2 Statistical properties of quarterly real GDP and GDP deflator

In order to avoid delays in data, data are gathered from original sources, in particular Philippines Statistical Authority and the Central Bank of the Philippines. Data are downloaded as Excel tables, and a database is created in year-on-year growth rates. Since GDP data are available from the first quarter of 1999 to the second quarter of 2018, monthly data for the same period were used in this chapter. Although, more recent monthly data may

be available, this avoids the issue of ragged-edge data, which is very important, but not the main focus of the chapter. Data are available in two Excel files (although it is possible to have them in the same Excel file with two sheets). Data may be downloaded with `read_excel` command (Wickham et al., 2018).

```
>library(readxl)
>url <- "https://web.sas.upenn.edu/ozmucur/files/2018/09/HOS410-
  1521dbn.xls"
>destfile <- "HOS410_1521dbn.xls"
>curl::curl_download(url, destfile)
>HOS410_1521dbn <- read_excel(destfile, col_types = c("date", "numeric", "numeric",
  "numeric", "numeric"))
>View(HOS410_1521dbn)

> cbind(tail(HOS410_1521dbn,3), head(HOS410_1521dbn,3))
      date      NOM      GDP      DEF      TIME      date      NOM      GDP      DEF
1 2017-10-01  8.648800  6.504828  2.013028  2017.75  1999-01-01  8.898427  0.259630  8.616427
2 2018-01-01  9.630288  6.575070  2.866729  2018.00  1999-04-01 10.100770  2.937917  6.958419
3 2018-04-01  9.563416  5.995075  3.366515  2018.25  1999-07-01  9.994815  4.020711  5.743187
      TIME
1 1999.00
2 1999.25
3 1999.50
```

3.2.1 Descriptive statistics for GDP and GDP deflator

Stats package provides minimum, maximum, first and third quartiles, mean and median (R Core Team, 2018). Real GDP growth from the first quarter of 1998 to the first quarter of 1999 was only 0.26%, while GDP deflator growth was 8.6%. On the other hand, the real GDP growth from the second quarter of 2017 to the second quarter of 2018 was 6.0%, while the GDP deflator growth was 3.4%. The mean for all 78 observations was 5.2% for real GDP, and 3.9% for GDP deflator.

```
>summary(HOS410_1521dbn)
      date                  NOM                  GDP                  DEF
Min.   :1999-01-01 00:00:00  Min.   :1.444   Min.   :0.2596  Min.   :-1.456
1st Qu.:2003-10-24 00:00:00  1st Qu.:8.516   1st Qu.:3.9845  1st Qu.: 2.456
Median :2008-08-16 00:00:00  Median :9.338   Median :5.6681  Median : 3.907
Mean   :2008-08-15 21:13:50  Mean   :9.261   Mean   :5.2082  Mean   : 3.866
3rd Qu.:2013-06-08 06:00:00  3rd Qu.:10.320  3rd Qu.:6.6317  3rd Qu.: 5.235
Max.   :2018-04-01 00:00:00  Max.   :15.403  Max.   :8.9131  Max.   : 9.665
      TIME
Min.   :1999
1st Qu.:2004
Median :2009
Mean   :2009
3rd Qu.:2013
```

Max. :2018

Kernel density estimates show that real GDP growth is skewed to the left (with a skewness coefficient of -0.59), and slightly less peaked than a normal distribution (with an excess kurtosis of -0.15). It has a mean of 5.2 with a standard deviation of 1.86 , median of 5.67 , and with a mode close to 7 . On the other hand, GDP deflator has a relatively close to a normal distribution (with a skewness coefficient of 0.15 , and an excess kurtosis 0.19). It has a mean of 3.87 , with a standard deviation of 2.1 .

Psych package ([Revelle, 2019](#)) describes command also provides, standard deviation (same as sd in Stats package) mean absolute deviation, trimmed mean, skewness, and kurtosis. Moments Package has tests for skewness and kurtosis.

```
>yts3<- yts[,2:4]
>round(describe(yts3),2)
   vars n mean   sd median trimmed   mad   min   max range skew kurtosis    se
NOM     1 78  9.26 2.30    9.34    9.31 1.38  1.44 15.40 13.96 -0.47    1.70 0.26
GDP     2 78  5.21 1.86    5.67    5.33 1.77  0.26  8.91  8.65 -0.59    -0.15 0.21
DEF     3 78  3.87 2.10    3.91    3.87 2.10 -1.46  9.66 11.12  0.05    0.19 0.24
```

3.2.2 Unit root tests for GDP and GDP deflator

Augmented Dickey–Fuller tests indicate that both GDP and GDP Deflator are stationary at the 1% level ([Hyndman et al., 2019](#); [Stoffer, 2017](#)). In these tests, the number of lags is selected based on Bayesian information criteria (BIC). Test statistic is -3.85 for real GDP, and -3.76 for GDP deflator, compared with the critical value of -3.51 at the 1% level.

```
>urootGDP<- ur.df(GDP,type="drift",selectlags="BIC")
>summary(urootGDP)

#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####

Test regression drift

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q      Max 
-2.4126 -0.5471  0.0946  0.5956  5.6054 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.65512   0.44185   3.746 0.000356 *** 
z.lag.1     -0.30973   0.08037  -3.854 0.000248 *** 
z.diff.lag   0.18314   0.10904   1.680 0.097306 .  

```

```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.188 on 73 degrees of freedom
Multiple R-squared: 0.1717, Adjusted R-squared: 0.149
F-statistic: 7.568 on 2 and 73 DF, p-value: 0.001031

Value of test-statistic is: -3.8538 7.4601

Critical values for test statistics:
    1pct  5pct 10pct
tau2 -3.51 -2.89 -2.58
phil  6.70  4.71  3.86

```

3.2.3 Autocorrelation and partial autocorrelation functions GDP and GDP deflator

Autocorrelation (ACF) and partial autocorrelation functions (PACF) can be used to check for stationarity and also to identify the order of an autoregressive integrated moving average (ARIMA) model.

Autocorrelation and partial autocorrelation coefficients for GDP show that only first and second order coefficients are significantly different from zero. Since coefficients do not necessarily decay, decrease with lags, one may have to use data after using a difference operator. A cursory look suggests that ARIMA(1,0,2) may be used as the initial model in Box–Jenkins methodology.

On the other hand, autocorrelation and partial autocorrelation coefficients for GDP Deflator (DEF) show that there are quite a few coefficients which are significantly different from zero. These include autocorrelations with lags 1 through 13, and partial autocorrelations of order one and two. In this case, ARIMA (2,0,3) or ARIMA (2,1,3) may be used as the initial model.

3.2.4 Correlations and cross-correlations: GDP and GDP deflator

A simple correlation matrix indicates that the correlation between GDP and Deflator is negative 0.39. Real GDP is positively related to time ($r = 0.50$), and the GDP Deflator is negatively related to time (-0.66). Both variables seem to have deterministic trends. Nominal GDP growth has a negative trend. Simple correlation gives the degree of the relationship of two variables at the same time period. It is important to find out how these correlations change with lags in one of the variable. Cross-correlations help to see how two variables are related with lags. For example, with both variables at zero lags, the correlation is negative 0.39. The correlation is higher if the GDP Deflator is lagged up to five periods (close to negative 0.6 at lags 3 and 4). These results suggest that, in general, the growth in GDP Deflator precedes real GDP growth with a relatively high negative correlation. Therefore, one can argue that a higher inflation (measured with GDP deflator) precedes a lower real GDP growth. Granger causality tests support this finding.

Granger causality tests can be done by estimating two equations for each variable: one with the lags of both variables (in this case GDP and DEF), and the other one with only the lagged values of the dependent variable (GDP). If the lagged values of DEF are statistically significantly different from zero, then DEF Granger causes GDP. Since, lagged values of DEF in the GDP equation are statistically significant as a group (F value of 9.96, P -value = 0.00015), DEF Granger causes GDP. On the other hand, lagged values of GDP in the DEF equation are not statistically significant (F = 0.386, P -value = 0.68), so GDP does not Granger cause DEF. Therefore, one can conclude that there is unidirectional causality from GDP Deflator growth (DEF) to real GDP growth (GDP) (similar results are obtained with four lags, not reported here).

```
> grangertest(DEF~GDP,order=2,na.action=na.omit)
Granger causality test

Model 1: DEF ~ Lags(DEF, 1:2) + Lags(GDP, 1:2)
Model 2: DEF ~ Lags(DEF, 1:2)
  Res.Df Df      F Pr(>F)
1     71
2     73 -2 0.386 0.6812

> grangertest(GDP~DEF,order=2,na.action=na.omit)
Granger causality test

Model 1: GDP ~ Lags(GDP, 1:2) + Lags(DEF, 1:2)
Model 2: GDP ~ Lags(GDP, 1:2)
  Res.Df Df      F Pr(>F)
1     71
2     73 -2 9.9654 0.0001533 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4 Estimated models in R

4.1 Box–Jenkins (ARIMA(p,d,q)) univariate time-series models

Based on autocorrelation and partial autocorrelation functions Box–Jenkins methodology suggests ARIMA(2,0,1) as the initial model for GDP. After estimating this model, diagnostic checking requires no serial correlation in residuals and statistically significant parameters. Once a model is chosen, it is also checked if a higher order model improves the fit. As a result of this process, ARIMA(2,1,2) is chosen as the best model. Residuals from this model are serially uncorrelated (Box–Ljung Q statistics), and stationary according to unit root tests. Autocorrelation and partial autocorrelation function of residuals also show that there are no systematic pattern in residuals.

Similar approach to DEF also suggests an ARIMA(2,1,2) model. Residuals from this model also have desired properties. However, there is the possible presence of a significant fourth order correlation, as indicated by the correlogram of residuals.

It is possible to determine the order of an ARIMA model with the help of an “auto.arima” command in the Forecast Package. This approach suggests an ARIMA(0,1,0) for real GDP (GDP) and ARIMA(2,1,1) for the GDP deflator (DEF). However, Akaike information criterion for models selected using Box–Jenkins methodology is lower than automated models. For example, GDP model selected by Box–Jenkins approach has an Akaike Information criterion of 252.64, compared with 261.84 using the automated approach (auto.arima). Similarly, DEF model selected by Box–Jenkins approach has an Akaike Information criterion of 224.74, compared with 225.47 using the automated approach (auto.arima).

```
> print(GDP_ARIMA212<- arima(GDP, order=c(2,1,2))) #chosen model
Call:
arima(x = GDP, order = c(2, 1, 2))

Coefficients:
      ar1      ar2      ma1      ma2
    1.6734 -0.8836 -1.8351  0.9454
  s.e.  0.0853  0.0674  0.1344  0.1396

sigma^2 estimated as 1.355: log likelihood = -122.32, aic = 252.64

> Box.test(residuals(GDP_ARIMA212), lag =74, type = "Ljung-Box",
fitdf=4)

        Box-Ljung test

data: residuals(GDP_ARIMA212)
X-squared = 40.888, df = 70, p-value = 0.9979

> acf(residuals(GDP_ARIMA212))
> ur.df(residuals(GDP_ARIMA212), type="none", selectlags="BIC")
#####
# Augmented Dickey-Fuller Test Unit Root / Cointegration Test #
#####

The value of the test statistic is: -6.5493
```

4.2 Vector autoregressive models

A vector autoregressive (VAR) model with two variables, GDP and DEF, is estimated as an alternative using the package VARS ([Pfaff and Stigler, 2018](#)).

First, the number of lags is determined using Akaike information criterion, Schwarz information criterion, Hannan–Quinn criterion, and forecast prediction error criterion. All information criteria suggest a model with two lags is the selected one. After selecting the number of lags, a VAR2 model is estimated. In this system, GDP equation has a determination coefficient of 0.66. Parameters associated with lagged values of GDP, GDP(t-1) and GDP(t-2) are both statistically significant at the 5% level. The parameter associated with DEF(t-2) is highly significant, but DEF(t-1) is not statistically different from zero. The DEF equation has a higher determination coefficient (0.76). In that equation, both lags of DEF are significant, but both lagged GDP variables are insignificant. The correlation between residuals from the GDP equation and the DEF equation is rather low (-0.10) indicating that contemporaneous variables (GDP, and DEF on the right hand side of the equations) are not probably needed in this system. The Granger causality test indicate a unidirectional causality from GDP deflator (DEF) to GDP. This is the same conclusion from a single equation Granger causality test. The test also rejects the instantaneous (contemporaneous) causality between GDP and DEF.

Both serial correlation and arch tests on residuals indicate that these issues are not statistically significant at the 5% level. Jarque–Bera test rejects the assumption of normality of residuals. Further analysis reveals that this is a result of kurtosis component, and not the skewness component.

The effect of a one standard deviation shock in GDP on GDP and DEF can be traced with the help of impulse response functions.

Variance decompositions reveal the effect of DEF on GDP. About a third of a forecast error in GDP can be explained DEF after six quarters. On the other hand, only 4% of forecast error in DEF can be explained by GDP.

```
> var12<- VARselect(dat1,12)
> var12
$'selection'
AIC(n)  HQ(n)  SC(n)  FPE(n)
2       2       2       2

$criteria
      1       2       3       4       5       6
AIC(n) 0.7595216 0.4189505 0.4760668 0.5121098 0.4670596 0.4474970
HQ(n)   0.8381795 0.5500469 0.6596019 0.7480834 0.7554718 0.7883478
SC(n)   0.9585812 0.7507163 0.9405390 1.1092884 1.1969445 1.3100883
FPE(n)  2.1375216 1.5212497 1.6123107 1.6745240 1.6053429 1.5807985
      7       8       9       10      11      12
AIC(n) 0.5196397 0.5818682 0.6931478 0.7250021 0.7077023 0.8155168
HQ(n)   0.9129291 1.0275962 1.1913143 1.2756072 1.3107460 1.4709991
SC(n)   1.5149373 1.7098722 1.9538581 2.1184187 2.2338253 2.4743462
FPE(n)  1.7088076 1.8323619 2.0681199 2.1614214 2.1568211 2.4471894
```

```

var2<- VAR(dat1,p=2)
> summary(var2)

VAR Estimation Results:
=====
Endogenous variables: GDP, DEF
Deterministic variables: const
Sample size: 76
Log Likelihood: -216.353
Roots of the characteristic polynomial:
0.6442 0.6442 0.5191 0.5191
Call:
VAR(y = dat1, p = 2)

Estimation results for equation GDP:
=====
GDP = GDP.11 + DEF.11 + GDP.12 + DEF.12 + const

      Estimate Std. Error t value Pr(>|t|)
GDP.11  0.75844   0.10548   7.190 5.25e-10 ***
DEF.11  0.15306   0.11271   1.358 0.178750
GDP.12 -0.22826   0.09823  -2.324 0.023009 *
DEF.12 -0.40285   0.11443  -3.520 0.000757 ***
const    3.47611   0.59148   5.877 1.24e-07 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.064 on 71 degrees of freedom
Multiple R-Squared: 0.6588, Adjusted R-squared: 0.6396
F-statistic: 34.28 on 4 and 71 DF, p-value: 6.409e-16

Estimation results for equation DEF:
=====
DEF = GDP.11 + DEF.11 + GDP.12 + DEF.12 + const

      Estimate Std. Error t value Pr(>|t|)
GDP.11 -0.08726   0.10108  -0.863 0.390854
DEF.11  1.17593   0.10800  10.888 < 2e-16 ***
GDP.12  0.04485   0.09413   0.476 0.635206
DEF.12 -0.41076   0.10966  -3.746 0.000362 ***
const    1.10555   0.56678   1.951 0.055054 .
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 71 degrees of freedom
Multiple R-Squared: 0.7602, Adjusted R-squared: 0.7467
F-statistic: 56.28 on 4 and 71 DF, p-value: < 2.2e-16

```

```

Covariance matrix of residuals:
      GDP      DEF
GDP  1.1331 -0.1118
DEF -0.1118  1.0404

Correlation matrix of residuals:
      GDP      DEF
GDP  1.000 -0.103
DEF -0.103  1.000

> causality(var2,cause="DEF")
$'Granger'

    Granger causality H0: DEF do not Granger-cause GDP

data: VAR object var2
F-Test = 9.9654, df1 = 2, df2 = 142, p-value = 8.914e-05

$Instant

    H0: No instantaneous causality between: DEF and GDP

data: VAR object var2
Chi-squared = 0.79773, df = 1, p-value = 0.3718

> causality(var2,cause="GDP")
$'Granger'

    Granger causality H0: GDP do not Granger-cause DEF

data: VAR object var2
F-Test = 0.386, df1 = 2, df2 = 142, p-value = 0.6805

$Instant

    H0: No instantaneous causality between: GDP and DEF

data: VAR object var2
Chi-squared = 0.79773, df = 1, p-value = 0.3718

> serial.test(var2)

    Portmanteau Test (asymptotic)

data: Residuals of VAR object var2
Chi-squared = 49.746, df = 56, p-value = 0.7091

> arch.test(var2)

    ARCH (multivariate)

data: Residuals of VAR object var2
Chi-squared = 58.799, df = 45, p-value = 0.08127

```

```

> normality.test(var2)
$'JB'

  JB-Test (multivariate)

data: Residuals of VAR object var2
Chi-squared = 23.55, df = 4, p-value = 9.83e-05

$Skewness

  Skewness only (multivariate)

data: Residuals of VAR object var2
Chi-squared = 1.0672, df = 2, p-value = 0.5865

$Kurtosis

  Kurtosis only (multivariate)

data: Residuals of VAR object var2
Chi-squared = 22.483, df = 2, p-value = 1.312e-05

```

4.3 Bridge equations

Bridge equations are used to determine the relationships between low-frequency variables (here quarterly real GDP and GDP deflator) and high-frequency indicators (here, monthly and monthly averages of daily indicators). These types of models were first proposed by [Klein and Sojo \(1987, 1989\)](#). Klein proposed to mimic the calculation of National Income and Product Accounts (NIPA) of the Bureau of Economic Analysis (BEA), when new data are available without waiting for the official release of NIPA. His insightful approach led to “the Current Quarter Model for the US Economy.” Some components of GDP are available monthly (for example, personal income and personal consumption expenditures). There is no reason to wait for the quarterly numbers to be released. One can use the monthly April personal income (released about the end of May) to estimate May and June figures and then the second quarter personal income (released about the end of July). There is a 2-month lag between the two.

Quarterly indicators are related to monthly indicators, one at a time; for example, as done in [Ghysels and Marcellino \(2018\)](#). Quarterly real GDP is related to quarterly averages of first 10 monthly indicators (X_{01} to X_{10}), and quarterly GDP deflator is related to monthly indicators (X_{11} to X_{17} , X_{09} and X_{10}).

For purposes of putting equations in a table format, the package Stargazer is particularly useful.

Four monthly indicators for real GDP have a better fit than others (coefficients of determination, R^2 's, ranging from 0.24 to 0.28). These indicators are

industrial production index (y-o-y growth) (X01), merchandise imports (y-o-y growth) (X02), real stock price index (y-o-y growth) (X07), and treasury bill rate (91-day)—US treasury bill rate (3-month) (X10).

Four monthly indicators for GDP deflator have a better fit than others (R^2 s ranging from 0.46 to 0.78) than others. These indicators are consumer price index (y-o-y growth) (X11), producer price index (y-o-y growth) (X12), wholesale price index, Metro Manila (y-o-y growth) (X13), and retail price index (y-o-y growth) (X14).

```
> summary(lm(GDP~X01))

Call:
lm(formula = GDP ~ X01)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8735 -0.9634  0.1137  1.0310  3.0974 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.8472    0.1770  27.391 < 2e-16 ***
X01         0.1016    0.0189   5.373 8.17e-07 ***  
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.523 on 76 degrees of freedom
Multiple R-squared: 0.2753, Adjusted R-squared: 0.2658 
F-statistic: 28.87 on 1 and 76 DF, p-value: 8.173e-07

> summary(lm(DEF~X11))

Call:
lm(formula = DEF ~ X11)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9111 -0.7307  0.1358  0.6082  1.9605 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.20563    0.24390   0.843   0.402    
X11         0.88909    0.05441  16.339 <2e-16 ***  
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.96 on 76 degrees of freedom
Multiple R-squared: 0.7784, Adjusted R-squared: 0.7755 
F-statistic: 267 on 1 and 76 DF, p-value: < 2.2e-16
```

4.4 Principal components using monthly indicators and bridge equations using principal components

Instead of using 10 indicators, the method of principal components is used to construct a weighted average of these indicators. These are based on 234 monthly indicators. There are two groups of monthly indicators. Indicators X01 to X10 are used to calculate principal components to be used in real GDP forecasts. Indicators X09 to X17 are used to calculate principal components in GDP deflator forecasts.

The first principal component is used. In real GDP, the first principal component explains 30% of variance, the first two components explain 59% of variance, and the first five components explain 77% of the variance.

The R^2 in the regression of real GDP on the quarterly average of the first monthly principal component (PC1) is 0.365. It is higher than that obtained in the regression of real GDP on individual indicators (X01 to X10). The second and the third principal components also have a statistically significant relationship with real GDP, but the coefficients of determination are not that high (0.06 for PC2 and 0.05 for PC3).

In the relationship with GDP deflator (DEF) and price related principal components, only the first (PC6) and the fourth (PC9) are significant. The coefficient of determination for PC6 is 0.80, and 0.06 for PC9. The determination coefficient of 0.80 is greater than the determination coefficient in the equation using X11.

```
> length(X01)Using Monthly Indicators
[1] 234
> data_Q<- H0S41M[1:10]
> library(psych)
> library(stats)
> principal(data_Q)
Principal Components Analysis
Call: principal(r = data_Q)
Standardized loadings (pattern matrix) based upon correlation matrix
    PC1     h2     u2   com
X01  0.58  0.341  0.66   1
X02  0.78  0.606  0.39   1
X03  0.63  0.392  0.61   1
X04 -0.17  0.029  0.97   1
X05  0.23  0.051  0.95   1
X06  0.72  0.512  0.49   1
X07  0.68  0.468  0.53   1
X08 -0.59  0.343  0.66   1
X09 -0.11  0.013  0.99   1
X10 -0.51  0.256  0.74   1
```

```

PC1
SS loadings    3.01
Proportion Var 0.30

Mean item complexity = 1
Test of the hypothesis that 1 component is sufficient.

The root mean square of the residuals (RMSR) is 0.14
with the empirical chi square 403.01 with prob < 4.1e-64

Fit based upon off diagonal values = 0.72
> PCforQ<-principal(data_Q)
> summary(PCforQ)

Factor analysis with Call: principal(r = data_Q)
Test of the hypothesis that 1 factor is sufficient.
The degrees of freedom for the model is 35 and the objective function was 1.23
The number of observations was 234 with Chi Square = 280.1 with prob < 5.2e-40

The root mean square of the residuals (RMSA) is 0.14
> PCforQ<-princomp(data_Q,cor=TRUE)
> summary(PCforQ)

Importance of components:
          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6
Standard deviation   1.7352437 1.2618927 1.1167697 1.0109934 0.93221761 0.79658812
Proportion of Variance 0.3011071 0.1592373 0.1247175 0.1022108 0.08690297 0.06345526
Cumulative Proportion 0.3011071 0.4603444 0.5850619 0.6872726 0.77417559 0.83763085
          Comp.7    Comp.8    Comp.9    Comp.10
Standard deviation   0.74701199 0.68749385 0.56123388 0.52728865
Proportion of Variance 0.05580269 0.04726478 0.03149835 0.02780333
Cumulative Proportion 0.89343354 0.94069832 0.97219667 1.00000000

> summary(lm(HOS41M$GDP~HOS41M$AP1))

Call:
lm(formula = HOS41M$GDP ~ HOS41M$AP1)

Residuals:
    Min      10 Median      30      Max 
-4.0943 -0.9651  0.2065  0.9765  2.5328 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.0615     0.1611   31.410 < 2e-16 ***
HOS41M$AP1  0.6365     0.0959    6.637 4.21e-09 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 1.423 on 76 degrees of freedom
Multiple R-squared: 0.3669, Adjusted R-squared: 0.3586
F-statistic: 44.05 on 1 and 76 DF, p-value: 4.208e-09

> summary(lm(HOS41M$DEF~HOS41M$AP6))

Call:
lm(formula = HOS41M$DEF ~ HOS41M$AP6)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.36252 -0.62037  0.03185  0.71214  2.10283 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 3.77313   0.10456  36.09 <2e-16 ***
HOS41M$AP6  0.94122   0.05483  17.17 <2e-16 ***  
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9235 on 76 degrees of freedom
Multiple R-squared: 0.795, Adjusted R-squared: 0.7923
F-statistic: 294.7 on 1 and 76 DF, p-value: < 2.2e-16

```

Bridge equations are not necessarily treated as single equation models. They can be used together to form a more complete model (Klein and Ozmucur, 2008; Klein & Park, 1993, 1995; Klein and Sojo, 1987). Here, there are two bridge equations: $GDP = f_1$ (PC1), $DEF = f_2$ (PC6). A third equation, nominal GDP growth, may complete the model: $NOM = GDP + DEF$. Obviously, this can be done in MIDAS framework also. One advantage of using quarterly averages of monthly variables (as in bridge equations) is to be able to generate ratio of variables and use them in an analysis. For example, if the share of money in nominal GDP is an important variable to use, it can be done by taking the ratio of quarterly averages of monthly money supply figures to nominal GDP. This may be even more desirable if principal components are constructed using quarterly data (which is done in Klein’s “Current Quarter Model”) and not monthly data.

4.5 MIDAS models

Three MIDAS regression models are used here: unrestricted MIDAS, exponential Almon weights, and beta distribution weights (Ghysels et al., 2016; Kvedaras and Zemlys, 2016). Six monthly observations (lags 0–5) are used. For example, second quarter real GDP is related to monthly variable for June, May, April, March, February, and January. Residual standard error for unrestricted MIDAS using X01 is 1.501, compared with 1.592 using bridge equation. It should be noted that MIDAS equation uses lags 0–5,

but bridge equation uses average of lags 0, 1, and 2 only. For a fair comparison, bridge equations should include average of lags 3, 4, and 5, also. However, since using the average of montly indicators in the relevant quarter is the more common use of bridge equation, we keep this comparison. Using exponential Almon weights, the residual standard error is 1.508, and using beta weights it is 1.506. There is not much difference between these three models.

Since using principal components improved bridge equations significantly, here we will concentrate on MIDAS models using principal components. To emphasize the difference in frequency, quarterly variables are renamed with a prefix of Q, and monthly variables are renamed with a prefix of M. The first principal component (PC1 in real GDP equations, and PC6 in GDP deflator equations) is used.

Unrestricted MIDAS has the lowest residual standard error among the three in real GDP equations. Residual standard errors are 1.393 in unrestricted model, 1.404 in exponential Almon, and 1.418 in beta distribution models. All of these MIDAS models have better fits compared to the bridge equation with principal components, which has the residual standard error is 1.491.

In the GDP deflator equations, residual standard errors are 0.961 in exponential Almon, 0.995 in beta, and 0.962 in unrestricted MIDAS, not much different from the residual standard error of 0.952 for the bridge equation model.

```
> beta0_GDP<- midas_r(Q_GDP~mls(M_PC1,0:5,3,nealmon),start=list(M_PC1=c(1,-0.5)))
> beta0_GDP
MIDAS regression model
model: Q_GDP ~ mls(M_PC1, 0:5, 3, nealmon)

(Intercept)      M_PC11      M_PC12
      5.2696     0.6983    -0.2217

Function optim was used for fitting
> summary(beta0_GDP)

Formula Q_GDP ~ mls(M_PC1, 0:5, 3, nealmon)
Parameters:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.2696     0.3896   13.526 < 2e-16 ***
M_PC11      0.6983     0.1469    4.752 9.64e-06 ***
M_PC12     -0.2217     0.4189   -0.529    0.598
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.404 on 74 degrees of freedom

> betan_GDP<-
midas_r(Q_GDP~mls(M_PC1,0:5,3,nbetaMT),start=list(M_PC1=c(2,1,5,0)))
```

```

> summary(betan_GDP)
Formula Q_GDP ~ mls(M_PC1, 0:5, 3, nbetaMT)
Parameters:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.278e+00 3.450e-01 15.297 < 2e-16 ***
M_PC11     6.627e-01 1.502e-01  4.411 3.54e-05 ***
M_PC12     1.804e+01 3.905e+03  0.005    0.996
M_PC13     1.800e+01 3.905e+03  0.005    0.996
M_PC14     2.201e-02 3.280e-01  0.067    0.947
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.418 on 72 degrees of freedom
> um_GDP<- midas_r(Q_GDP~ mls(M_PC1, 0:5, 3), start = NULL)
> summary(um_GDP)
Formula Q_GDP ~ mls(M_PC1, 0:5, 3)
Parameters:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.31395    0.32025 16.593 < 2e-16 ***
M_PC11     0.58600    0.28333  2.068  0.04231 *
M_PC12    -0.47259    0.32699 -1.445  0.15285
M_PC13     0.17938    0.23997  0.747  0.45727
M_PC14     0.64712    0.23327  2.774  0.00709 **
M_PC15    -0.12033    0.32353 -0.372  0.71106
M_PC16    -0.08815    0.18077 -0.488  0.62734
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.393 on 70 degrees of freedom.

```

4.6 ADL-MIDAS

The MIDAS model can be extended to include autoregressive terms ([Ghysels and Marcellino, 2018](#)). Models with one and two lags are given here. Although, it is theoretically feasible, there may be some complications related to nonlinear estimation (an example here is the DEF equation with beta weights for monthly indicators and with two autoregressive lags). Comparison with MIDAS regression results indicate an improvement in fit because of autoregressive lags. For example, unrestricted U-MIDAS for GDP has a residual standard error of 1.39. The standard error for ADL-MIDAS of the same form is 1.03 for one or two autoregressive lags. There is not much difference between one or two autoregressive lags. Therefore, one lag may be sufficient. Similar results are obtained for DEF. For example, U-MIDAS for DEF equation has a residual standard error of 0.96,

compared with 0.69 for ADL-MIDAS. These results suggest that ADL-MIDAS may provide some improvement in model fit. It is still necessary to see if it also improves the forecasting ability of the model.

```

> beta0_GDP1<- midas_r(Q_GDP~ mls(Q_GDP, 1:1,
1)+mls(M_PC1,0:5,3,nealmon),start=list(M_PC1=c(1,-0.5)))
> beta0_GDP1
MIDAS regression model
model: Q_GDP ~ mls(Q_GDP, 1:1, 1) + mls(M_PC1, 0:5, 3, nealmon)
(Intercept)      Q_GDP       M_PC11      M_PC12
    2.2858      0.5746      0.3812     -1.6957

Function optim was used for fitting
> summary(beta0_GDP1)

Formula Q_GDP ~ mls(Q_GDP, 1:1, 1) + mls(M_PC1, 0:5, 3, nealmon)

Parameters:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.28579   0.56747   4.028 0.000136 ***
Q_GDP        0.57459   0.09652   5.953 8.46e-08 ***
M_PC11       0.38118   0.13277   2.871 0.005352 **
M_PC12      -1.69566   3.19231  -0.531 0.596913
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.063 on 73 degrees of freedom

> betan_GDP1<- midas_r(Q_GDP~ mls(Q_GDP, 1:1,
1)+mls(M_PC1,0:5,3,nbetaMT),start=list(M_PC1=c(2,1,5,0)))
> summary(betan_GDP1)

Formula Q_GDP ~ mls(Q_GDP, 1:1, 1) + mls(M_PC1, 0:5, 3, nbetaMT)

Parameters:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.93605   0.42167   4.591 1.86e-05 ***
Q_GDP        0.64085   0.07419   8.638 1.10e-12 ***
M_PC11       0.29335   0.15128   1.939   0.0565 .
M_PC12       1.00082   0.01823  54.909 < 2e-16 ***
M_PC13       1.09054   0.52826   2.064   0.0426 *
M_PC14      -0.14878   0.01766  -8.426 2.71e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.022 on 71 degrees of freedom

```

```

> um_GDP1<- midas_r(Q_GDP~ mls(Q_GDP, 1, 1)+ mls(M_PC1, 0:5, 3),
+ start = NULL)
> summary(um_GDP1)

Formula Q_GDP ~ mls(Q_GDP, 1:1, 1) + mls(M_PC1, 0:5, 3)

Parameters:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.01070 0.50308 3.997 0.000159 ***
Q_GDP       0.62872 0.08663 7.258 4.53e-10 ***
M_PC11      0.24864 0.19366 1.284 0.203470
M_PC12      -0.02829 0.29588 -0.096 0.924113
M_PC13      0.22790 0.19933 1.143 0.256864
M_PC14      0.24159 0.24135 1.001 0.320312
M_PC15      -0.03609 0.25395 -0.142 0.887406
M_PC16      -0.35487 0.14708 -2.413 0.018491 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.032 on 69 degrees of freedom

```

4.7 MIDAS-VAR

Ghysels (2016b) introduces a mixed-frequency VAR representation, in which high and low frequency data are stacked as skip-sampled processes. This observation-driven approach, called MIDAS-VAR, includes all quarterly variables (GDP and DEF) and corresponding monthly indicators (here, the first principal component (PC1) for GDP, and the first principal component (PC6) for DEF) with zero, one, and two lags as variables (PC1_LAG0, PC1_LAG1, PC1_LAG2, PC6_LAG0, PC6_LAG1, and PC6_LAG2) in a VAR system. Standard VAR techniques can be applied to this model with eight variables (Ghysels and Marcellino, 2018). Since, the number of parameters to be estimated can increase very fast with this model, a maximum of four lags is allowed. Selection criteria suggest one (HQ and SC) or four (AIC, and FPE) lags. Here, one lag is used. MIDAS-VAR improves the fit when compared with VAR. For example, residual standard error for GDP is 1.003 for MIDAS-VAR2, compared with residual standard error of 1.015 in standard VAR2. However, there is no improvement in GDP if MIDAS-VAR of order one is used. There is improvement in DEF estimation. MIDAS-VAR2 has residual standard error of 0.8368, compared with 0.9832 in standard VAR2 model. The residual standard error for DEF equation is 0.8566 for MIDAS-VAR1. Granger causality tests indicate that Granger type causality from DEF to GDP is not supported in this model. However, there are two major conclusions can be drawn. First, the null hypothesis of no instantaneous causality is

rejected. This can also be seen from high correlation coefficients among residuals from different equations. Second, there is Granger causality from monthly principal components (PC1 and PC6) with two lags to GDP and DEF.

```
>testlag<- VARselect(midasvar1,lag.max=4)
>testlag
$`selection`
AIC(n)  HQ(n)  SC(n)  FPE(n)
4        1        1        4

$criteria
      1           2           3           4
AIC(n) -9.528587e+00 -9.610049e+00 -9.862554e+00 -1.059201e+01
HQ(n)  -8.634308e+00 -7.920856e+00 -7.378446e+00 -7.312986e+00
SC(n)  -7.286794e+00 -5.375551e+00 -3.635351e+00 -2.372100e+00
FPE(n) 7.344997e-05  7.168285e-05  6.497142e-05  4.305254e-05

> MIDASVAR1<- VAR(midasvar1,p=1)
> summary(MIDASVAR1)

VAR Estimation Results:
=====
Endogenous variables: GDP, DEF, PC1_LAGO, PC1_LAG1, PC1_LAG2,
PC6_LAGO, PC6_LAG1, PC6_LAG2
Deterministic variables: const
Sample size: 77
Log Likelihood: -461.095
Roots of the characteristic polynomial:
0.7474 0.7166 0.7166 0.5237 0.4126 0.3164 0.3164 0.1444
Call:
VAR(y = midasvar1, p = 1)
```

Estimation results for equation GDP:

=====

```
GDP = GDP.11 + DEF.11 + PC1_LAGO.11 + PC1_LAG1.11 + PC1_LAG2.11
+ PC6_LAGO.11 + PC6_LAG1.11 + PC6_LAG2.11 + const
          Estimate Std. Error t value Pr(>|t|)
GDP.11      0.50147   0.09728   5.155 2.37e-06 ***
DEF.11     -0.18293   0.14356  -1.274 0.206903
PC1_LAGO.11  0.41996   0.22703   1.850 0.068688 .
PC1_LAG1.11  0.17038   0.28254   0.603 0.548484
PC1_LAG2.11 -0.30597   0.17433  -1.755 0.083738 .
PC6_LAGO.11 -0.09505   0.36397  -0.261 0.794770
PC6_LAG1.11  0.43915   0.58243   0.754 0.453449
PC6_LAG2.11 -0.30729   0.32729  -0.939 0.351105
const       3.29016   0.86443   3.806 0.000305 ***
---
```

```
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 1.038 on 68 degrees of freedom

Multiple R-Squared: 0.6667, Adjusted R-squared: 0.6275

F-statistic: 17 on 8 and 68 DF, p-value: 1.438e-13

Estimation results for equation DEF:

```
=====
DEF = GDP.11 + DEF.11 + PC1_LAGO.11 + PC1_LAG1.11 + PC1_LAG2.11
+ PC6_LAGO.11 + PC6_LAG1.11 + PC6_LAG2.11 + const
      Estimate Std. Error t value Pr(>|t|)
GDP.11     -0.11529   0.08024 -1.437  0.15538
DEF.11      0.67900   0.11841  5.734 2.46e-07 ***
PC1_LAGO.11 0.22858   0.18726  1.221  0.22644
PC1_LAG1.11 -0.21744   0.23305 -0.933  0.35410
PC1_LAG2.11  0.10572   0.14379  0.735  0.46474
PC6_LAGO.11  1.55934   0.30022  5.194 2.04e-06 ***
PC6_LAG1.11 -1.47943   0.48041 -3.080  0.00299 **
PC6_LAG2.11  0.08336   0.26996  0.309  0.75844
const        1.81509   0.71301  2.546  0.01318 *
---
```

```
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.8561 on 68 degrees of freedom.

Multiple R-Squared: 0.8314, Adjusted R-squared: 0.8116

F-statistic: 41.92 on 8 and 68 DF, P-value: <2.2e-16.

4.8 VARX and VARXM

Here, two extensions of VAR model are considered. The VARX model adds quarterly average of first monthly principal components (monthly averages PC1 and PC6) as exogenous variables in a VAR system. These exogeneous variables are called AP1, and AP6. Each equation in this system may be regarded as a bridge equation with lagged dependent variables. On the other hand, VARXM model adds first 6 monthly obervations as exogenous variables. Each equation resembles an ADL-MIDAS equation.

Both AP1 and AP6 improve the fit in the system. The residual standard error in GDP equation is 0.943 in VAR2X, compared with 1.015 in VAR2. On the other hand, the residual standard error in DEF equation is 0.748 in VAR2X, compared with 0.983 in VAR2. There is unidirectional Granger causality from DEF to GDP. There is no instantenous causality between DEF and GDP at the 5% level.

The improvement in the fit is even better if VARXM model is used. The residual standard error in GDP equation is 0.908 in VAR2XM, compared with 1.015 in VAR2, and the residual standard error in DEF equation is 0.673 in VAR2XM, compared with 0.983 in VAR2.

Both VARX and VARXM perform better, have lower residual standard errors, when compared with VAR and MIDAS-VAR. Correlations in residuals from different equations in MIDAS-VAR were pointing toward a VARXM model.

```
> exogen<-dat[,3:4]
> VAR2X<- VAR(dat1,p=2,exogen=exogen)
> summary(VAR2X)
```

VAR Estimation Results:

```
=====
Endogenous variables: GDP, DEF
Deterministic variables: const
Sample size: 76
Log Likelihood: -179.806
Roots of the characteristic polynomial:
0.4748 0.4573 0.4573 0.2445
Call:
VAR(y = dat1, p = 2, exogen = exogen)
```

Estimation results for equation GDP:

```
=====
GDP = GDP.11 + DEF.11 + GDP.12 + DEF.12 + const + AP1 + AP6
```

| | Estimate | Std. Error | t value | Pr(> t) |
|--------|----------|------------|---------|--------------|
| GDP.11 | 0.62638 | 0.10467 | 5.984 | 8.70e-08 *** |
| DEF.11 | 0.12907 | 0.15260 | 0.846 | 0.400578 |
| GDP.12 | -0.17613 | 0.09255 | -1.903 | 0.061190 . |
| DEF.12 | -0.29597 | 0.11833 | -2.501 | 0.014755 * |
| const | 3.47034 | 0.60137 | 5.771 | 2.05e-07 *** |
| AP1 | 0.27131 | 0.07661 | 3.542 | 0.000718 *** |
| AP6 | -0.05830 | 0.10242 | -0.569 | 0.571032 |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9434 on 69 degrees of freedom

Multiple R-Squared: 0.7141, Adjusted R-squared: 0.6892

F-statistic: 28.72 on 6 and 69 DF, p-value: < 2.2e-16

Estimation results for equation DEF:

```
=====
DEF = GDP.11 + DEF.11 + GDP.12 + DEF.12 + const + AP1 + AP6

      Estimate Std. Error t value Pr(>|t|)
GDP.11 -0.11087   0.08294 -1.337  0.18570
DEF.11  0.53005   0.12091  4.384 4.08e-05 ***
GDP.12  0.01228   0.07333  0.168  0.86746
DEF.12 -0.11716   0.09376 -1.250  0.21569
const    2.70278   0.47650  5.672 3.04e-07 ***
AP1     0.18352   0.06070  3.023  0.00351 **
AP6     0.55830   0.08115  6.880 2.20e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7475 on 69 degrees of freedom
Multiple R-Squared:  0.8654,    Adjusted R-squared:  0.8537
F-statistic: 73.95 on 6 and 69 DF,    p-value: < 2.2e-16
```

4.9 ARIMA and MIDAS

ADL-MIDAS and MIDAS-VAR use mixed frequencies to improve the model. They include autoregressive components, lagged values of quarterly or monthly variables. Is it possible to make use of the moving average components? This may be done in a two-step procedure. As the first step, an ARIMA (p, d, q) is estimated. In the second step, one looks for the relationship between residuals from this equation and high-frequency indicators, in our case monthly principal components. In the final step, predictions from the second equation are added to predictions from the first equation. A univariate model is improved by the use of high-frequency indicators.

```
> GDP_arima<- arima(GDP, order=c(2,1,2))

> REG1<-
  lm(residuals(GDP_arima)~PC1_LAG0+PC1_LAG1+PC1_LAG2+PC1_LAG3+PC1_LAG4+PC1_LAG5)
> summary(REG1)

Call:
lm(formula = residuals(GDP_arima) ~ PC1_LAG0 + PC1_LAG1 +
PC1_LAG2 +
PC1_LAG3 + PC1_LAG4 + PC1_LAG5)

Residuals:
    Min      1Q  Median      3Q      Max 
-2.3898 -0.5548  0.0090  0.6101  3.6712
```

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|----------------|-----------|------------|----------|----------|
| (Intercept) | 0.131032 | 0.121230 | 1.081 | 0.2835 |
| PC1_LAG0 | -0.022712 | 0.231469 | -0.098 | 0.9221 |
| PC1_LAG1 | 0.202755 | 0.284644 | 0.712 | 0.4786 |
| PC1_LAG2 | 0.205348 | 0.209146 | 0.982 | 0.3296 |
| PC1_LAG3 | 0.001862 | 0.245587 | 0.008 | 0.9940 |
| PC1_LAG4 | 0.024901 | 0.279512 | 0.089 | 0.9293 |
| PC1_LAG5 | -0.314838 | 0.168326 | -1.870 | 0.0656 . |
| --- | | | | |
| Signif. codes: | 0 ‘***’ | 0.001 ‘**’ | 0.01 ‘*’ | 0.05 ‘.’ |
| | 0.1 ‘ ’ | 1 | | |

Residual standard error: 1.053 on 70 degrees of freedom.

(1 observation deleted due to missing data)

Multiple R-squared: 0.1753, Adjusted R-squared: 0.1046

F-statistic: 2.48 on 6 and 70 DF, P-value: 0.03112.

4.10 Factor MIDAS

The method described here is a two-step procedure: MIDAS with explanatory variables in the form of factors that are first estimated from a factor analysis of the monthly indicators.

Factor Analysis Package ([Gilbert and Meijer, 2015](#)) is used with 15 monthly indicators (exchange rate (y-o-y growth) (X15), and money supply (M1) (y-o-y growth) (X16) are excluded from the list of 17 because they also appear in real terms. These 15 monthly variables are assumed to be determined by four factors (unobserved). These four factors explain close to 60% (0.587) of the variance. The first factor explains 21% of the variance, the second factor 16%, the third factor 12%, and the fourth factor 10% of the variance. These factors are extracted and used in a MIDAS model. This method is similar to using principal components, but actually very different because in a factor model unobserved variables (factors) are on the right hand side of the equation ($X = AF + v$), while in principal components analysis, unobserved variables (principal components) are on the left hand side of the equation ($P = BX + u$). In these equations, X are observed, F and P are unobserved, and v and u are stochastic disturbance terms. Monthly factor scores are extracted and named as FA4_1, FA4_2, FA4_3, and FA4_4. Quarterly real GDP growth is related to these factors one by one in an unrestricted MIDAS (U-MIDAS) framework with lags 1 to 6 (for example, FA4_11....FA4_16).

Since results are similar to the ones using principal components with two groups, here just the unrestricted MIDAS findings are given. Results suggest that the first factor (FA4_1) is related to prices and the second factor (FA4_2) is related to the real side of the economy.

```

> m2<- cbind(X01,X02,X03,X04,X05,X06,X07,X08,X09,X10,X11,X12,X13,X14,X17)
> factanal(m2,factors=4)

Call:
factanal(x = m2, factors = 4)

Uniquenesses:
   X01    X02    X03    X04    X05    X06    X07    X08    X09    X10    X11    X12    X13
X14    X17
  0.433 0.335 0.771 0.931 0.668 0.157 0.441 0.005 0.851 0.613 0.041 0.005 0.145
  0.254 0.552

Loadings:
          Factor1 Factor2 Factor3 Factor4
X01           0.600 -0.447
X02  -0.102   0.793 -0.160
X03           0.451 -0.155
X04  -0.156  -0.116 -0.106  0.140
X05  -0.538           -0.194
X06   0.251   0.788  0.317 -0.242
X07  -0.320   0.349 -0.572
X08  -0.213  -0.149  0.131  0.954
X09  -0.115  -0.116 -0.349
X10   0.305  -0.236  0.369  0.320
X11   0.924  -0.144  0.290
X12   0.351   0.915  0.182
X13   0.680   0.500  0.378
X14   0.844   0.175
X17   0.642   0.180

          Factor1 Factor2 Factor3 Factor4
SS loadings     3.209   2.347   1.729   1.514
Proportion Var  0.214   0.156   0.115   0.101
Cumulative Var  0.214   0.370   0.486   0.587

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 236.54 on 51 degrees of freedom.
The p-value is 1.08e-25

> FA4<- factanal(m2,factors=4,scores="regression",rotation="varimax")
> FA4_1<- factanal(m2,factors=4,scores="regression",rotation="varimax")$scores[,1]
> FA4_2<- factanal(m2,factors=4,scores="regression",rotation="varimax")$scores[,2]
> FA4_3<- factanal(m2,factors=4,scores="regression",rotation="varimax")$scores[,3]
> FA4_4<- factanal(m2,factors=4,scores="regression",rotation="varimax")$scores[,4]

> library(midasr)
> library(readxl)
> HOS41Q <- read_excel("HOS41Q.xls")

```

```

> View(HOS41Q)
> GDP<- HOS41Q$GDP
> Q_GDP<- ts(GDP,start=c(1999,1),end=c(2018,2),frequency=4)
> um_GDPxx2<- midas_r(Q_GDP ~ mls(FA4_2, 1:6, 3), start = NULL)
> summary(um_GDPxx2)

Formula Q_GDP ~ mls(FA4_2, 1:6, 3)

Parameters:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.3083     0.3887 13.655 <2e-16 ***
FA4_21    0.1923     0.4090  0.470  0.6396
FA4_22    0.8515     0.6562  1.298  0.1987
FA4_23    1.3123     0.7644  1.717  0.0905 .
FA4_24   -0.9433     0.6372 -1.481  0.1433
FA4_25   -1.1982     0.7146 -1.677  0.0981 .
FA4_26    1.1453     0.4581  2.500  0.0148 *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.434 on 69 degrees of freedom

> Q_DEF<- (DEF,start=c(1999,1),end=c(2018,2),frequency=4)
> um_DEFxx1<- midas_r(Q_DEF ~ mls(FA4_1, 1:6, 3), start = NULL)
> summary(um_DEFxx1)

Formula Q_DEF ~ mls(FA4_1, 1:6, 3)

Parameters:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.8471     0.3199 12.025 <2e-16 ***
FA4_11    1.5706     0.5980  2.626  0.0106 *
FA4_12   -0.1507     0.8565 -0.176  0.8609
FA4_13    1.5862     1.0567  1.501  0.1379
FA4_14   -1.8367     1.0520 -1.746  0.0853 .
FA4_15    0.1514     0.7855  0.193  0.8477
FA4_16    0.5237     0.5977  0.876  0.3840
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.215 on 69 degrees of freedom

```

4.11 Dynamic factor models—Small number of indicators

If the number of indicators is small, a state space framework may be used with all quarterly and monthly indicators (Mariano and Murasawa, 2003). Here, a similar example is given. There are several packages with state space models,

including Stats (R Statistical Functions, by [R Core Team \(2018\)](#)), Astsa (Applied Statistical Time Series Analysis, by [Stoffer \(2017\)](#)), DLM (Bayesian and Likelihood Analysis of Dynamic Linear Models, by [Petris \(2018\)](#)), KFAS (Kalman Filter and Smoother for Exponential Family State Space Models, by [Helske \(2018\)](#)), FKF (Fast Kalman Filter, by [Lue thi et al. \(2018\)](#)), and MARSS (Multivariate Autoregressive State-Space Modeling) by [Holmes et al. \(2018\)](#)). Here, MARSS is used. In this example, with one state process. There are 2 quarterly and 17 monthly variables in the system. There are 234 monthly observations (January 1999–June 2018). As in [Mariano and Murasawa \(2003\)](#), quarterly variables appear on the last month of the quarter. Kalman filter allows estimation of missing values.

A full MARSS model, with Gaussian errors, takes the form:

State equation:

$$X(t) = B(t)X(t-1) + u(t) + C(t)c(t) + G(t)w(t), \text{ where } w(t) \sim MVN(0, Q(t))$$

Measurement equation:

$$Y(t) = Z(t)X(t) + a(t) + D(t)d(t) + H(t)v(t), \text{ where } v(t) \sim MVN(0, R(t))$$

Initial Conditions:

$$X(1) \sim MVN(p, A) \text{ or } X(0) \sim MVN(p, A)$$

In a dynamic factor model framework, B matrix is an Identity matrix. EM algorithm is used in estimation.

```
> print(dfa4,what="model")
Model form is dfa. Model Structure is
m: 1 state process(es) named X1
n: 19 observation time series named GDP DEF X01 X02 X03 X04 X05
X06 X07 X08 X09 X10 X11 X12 X13 X14 X15 X16 X17

Z : unconstrained (19 x 1)
A : fixed and zero (19 x 1)
R : diagonal and equal (19 x 19)
B : fixed and all one (1 x 1)
U : fixed and zero (1 x 1)
Q : fixed and all one (1 x 1)
x0 : fixed and zero (1 x 1)
V0 : fixed and all 5 (1 x 1)
D : fixed and zero (19 x 1)
C : fixed and zero (1 x 1)
d : fixed and zero (1 x 1)
c : fixed and zero (1 x 1)
G : fixed and all one (1 x 1)
H : identity (19 x 19)
L : fixed and all one (1 x 1)
```

```
> dfa4<- MARSS(y,model=list(),form="dfa")
Success! abstol and log-log tests passed at 64 iterations.
Alert: conv.test.slope.tol is 0.5.
Test with smaller values (<0.1) to ensure convergence.

MARSS fit is
Estimation method: kem
Convergence test: conv.test.slope.tol = 0.5, abstol = 0.001
Estimation converged in 64 iterations.
Log-likelihood: -5327.154
AIC: 10694.31 AICc: 10694.51

      Estimate
Z.11   -0.12001
Z.21    0.25067
Z.31   -0.06314
Z.41   -0.03190
Z.51   -0.00603
Z.61   -0.04349
Z.71   -0.20350
Z.81    0.08894
Z.91   -0.12532
Z.101   0.01145
Z.111   -0.08487
Z.121   0.17015
Z.131   0.24958
Z.141   0.22233
Z.151   0.21469
Z.161   0.22380
Z.171   0.04736
Z.181   -0.14378
Z.191   0.15794
R.diag  0.74024
Initial states (x0) defined at t=0

Standard errors have not been calculated.
Use MARSSparamCIs to compute CIs and bias estimates.

> MARSSparamCIs(dfa4)

MARSS fit is
Estimation method: kem
Convergence test: conv.test.slope.tol = 0.5, abstol = 0.001
Estimation converged in 64 iterations.
Log-likelihood: -5327.154
AIC: 10694.31 AICc: 10694.51
```

| | ML.Est | Std.Err | low.CI | up.CI |
|------------------------------------|----------|---------|---------|-----------|
| Z.11 | -0.12001 | 0.0300 | -0.1787 | -0.061275 |
| Z.21 | 0.25067 | 0.0380 | 0.1763 | 0.325056 |
| Z.31 | -0.06314 | 0.0170 | -0.0965 | -0.029791 |
| Z.41 | -0.03190 | 0.0158 | -0.0630 | -0.000851 |
| Z.51 | -0.00603 | 0.0155 | -0.0363 | 0.024258 |
| Z.61 | -0.04349 | 0.0159 | -0.0747 | -0.012243 |
| Z.71 | -0.20350 | 0.0262 | -0.2548 | -0.152145 |
| Z.81 | 0.08894 | 0.0173 | 0.0550 | 0.122922 |
| Z.91 | -0.12532 | 0.0202 | -0.1649 | -0.085781 |
| Z.101 | 0.01145 | 0.0152 | -0.0184 | 0.041301 |
| Z.111 | -0.08487 | 0.0176 | -0.1194 | -0.050370 |
| Z.121 | 0.17015 | 0.0236 | 0.1239 | 0.216415 |
| Z.131 | 0.24958 | 0.0305 | 0.1898 | 0.309401 |
| Z.141 | 0.22233 | 0.0282 | 0.1670 | 0.277616 |
| Z.151 | 0.21469 | 0.0268 | 0.1622 | 0.267228 |
| Z.161 | 0.22380 | 0.0278 | 0.1694 | 0.278245 |
| Z.171 | 0.04736 | 0.0161 | 0.0158 | 0.078882 |
| Z.181 | -0.14378 | 0.0213 | -0.1856 | -0.102003 |
| Z.191 | 0.15794 | 0.0222 | 0.1144 | 0.201466 |
| R.diag | 0.74024 | 0.0166 | 0.7078 | 0.772733 |
| Initial states (x0) defined at t=0 | | | | |

CIs calculated at alpha = 0.05 via method=hessian

4.12 Dynamic factor models—Large number of indicators

Doz et al. (2011) use factor analysis to reduce the number of variables from a large number to a few. Here, for expository purposes, four factors were extracted from 15 monthly indicators. In general, the number of indicators to be included in this analysis is much larger. The method involves three steps. The first step is to use factor analysis to reduce the number of indicators (similar to factor MIDAS given in Section 4.10). The second step is to use these factors in a VAR framework. In the third step, predicted values of factors from the VAR model is used to forecast the target variables (GDP and DEF), after taking the quarterly averages of monthly factors (similar to a bridge equation). The model without the VAR is called the “static factor model,” with VAR it is called the “dynamic factor model.” The third step involves the use of Kalman filter to allow for varying coefficients. Here, the KFAS package by Helske (2018) is used. Lagged values of quarterly variables or quarterly averages of monthly variables may be used, rather than contemporaneous.

```
> varhazir<- cbind(FA4_1,FA4_2,FA4_3,FA4_4)
> factor4<- VAR(varhazir,p=2)
> factor477<- VARselect(varhazir,lag.max=12)
> factor477
```

```
$`selection`
AIC(n)  HQ(n)  SC(n) FPE(n)
    7      2      2      7

$criteria
          1           2           3           4           5           6
AIC(n) -9.978598e+00 -1.035309e+01 -1.032390e+01 -1.032150e+01 -1.031836e+01 -1.030751e+01
HQ(n)  -9.854833e+00 -1.013031e+01 -1.000211e+01 -9.900704e+00 -9.798544e+00 -9.688687e+00
SC(n)  -9.672050e+00 -9.801301e+00 -9.526872e+00 -9.279243e+00 -9.030857e+00 -8.774773e+00
FPE(n) 4.638348e-05  3.189987e-05  3.285656e-05  3.295715e-05  3.309637e-05  3.350978e-05
          7           8           9           10          11          12
AIC(n) -1.038162e+01 -10.344225000 -1.032666e+01 -1.032729e+01 -1.024592e+01 -1.030156e+01
HQ(n)  -9.663779e+00 -9.527377317 -9.410801e+00 -9.312421e+00 -9.132041e+00 -9.088663e+00
SC(n)  -8.603640e+00 -8.321011421 -8.058209e+00 -7.813603e+00 -7.486996e+00 -7.297392e+00
FPE(n) 3.118388e-05   0.000032465  3.316203e-05  3.329372e-05  3.631972e-05  3.458762e-05

> factor4<- VAR(varhazir,p=2)

> summary(factor4)

VAR Estimation Result:
=====
Endogenous variables: FA4_1, FA4_2, FA4_3, FA4_4
Deterministic variables: const
Sample size: 232
Log Likelihood: -98.664
Roo of the characteristic polynomial:
0.9303 0.9184 0.8832 0.8832 0.6175 0.3811 0.1695 0.0194
Call:
VAR(y = varhazir, p = 2)

Estimation result for equation FA4_1:
=====
FA4_1 = FA4_1.11 + FA4_2.11 + FA4_3.11 + FA4_4.11 + FA4_1.12 + FA4_2.12 +
FA4_3.12 + FA4_4.12 + const

      Estimate Std. Error t value Pr(>|t|)    
FA4_1.11  1.409893  0.071981 19.587 < 2e-16 ***
FA4_2.11  0.104278  0.046732  2.231 0.026649 *  
FA4_3.11  0.203644  0.060974  3.340 0.000983 *** 
FA4_4.11  0.077107  0.054357  1.419 0.157431    
FA4_1.12 -0.479108  0.070954 -6.752 1.24e-10 ***
FA4_2.12 -0.074202  0.045213 -1.641 0.102172    
FA4_3.12 -0.173690  0.060439 -2.874 0.004448 ** 
FA4_4.12 -0.097238  0.054978 -1.769 0.078316 .  
const     -0.005946  0.015630 -0.380 0.704023    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.2375 on 223 degrees of freedom
 Multiple R-Squared: 0.9376, Adjusted R-squared: 0.9354
 F-statistic: 419 on 8 and 223 DF, p-value: < 2.2e-16

Estimation resul for equation FA4_2:

```
=====
FA4_2 = FA4_1.11 + FA4_2.11 + FA4_3.11 + FA4_4.11 + FA4_1.12 + FA4_2.12 +
FA4_3.12 + FA4_4.12 + const
```

| | Estimate | Std. Error | t value | Pr(> t) |
|----------|----------|------------|---------|--------------|
| FA4_1.11 | 0.35174 | 0.10574 | 3.326 | 0.00103 ** |
| FA4_2.11 | 0.81134 | 0.06865 | 11.819 | < 2e-16 *** |
| FA4_3.11 | 0.20632 | 0.08957 | 2.303 | 0.02217 * |
| FA4_4.11 | -0.04398 | 0.07985 | -0.551 | 0.58235 |
| FA4_1.12 | -0.43680 | 0.10423 | -4.191 | 4.01e-05 *** |
| FA4_2.12 | 0.05823 | 0.06642 | 0.877 | 0.38161 |
| FA4_3.12 | -0.22729 | 0.08878 | -2.560 | 0.01113 * |
| FA4_4.12 | -0.02331 | 0.08076 | -0.289 | 0.77314 |
| const | 0.01507 | 0.02296 | 0.656 | 0.51230 |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3488 on 223 degrees of freedom
 Multiple R-Squared: 0.8677, Adjusted R-squared: 0.8629
 F-statistic: 182.8 on 8 and 223 DF, p-value: < 2.2e-16

Estimation resul for equation FA4_3:

```
=====
FA4_3 = FA4_1.11 + FA4_2.11 + FA4_3.11 + FA4_4.11 + FA4_1.12 + FA4_2.12 +
FA4_3.12 + FA4_4.12 + const
```

| | Estimate | Std. Error | t value | Pr(> t) |
|----------|------------|------------|---------|------------|
| FA4_1.11 | -0.0450646 | 0.0948683 | -0.475 | 0.635 |
| FA4_2.11 | 0.0472294 | 0.0615911 | 0.767 | 0.444 |
| FA4_3.11 | 0.8783935 | 0.0803622 | 10.930 | <2e-16 *** |
| FA4_4.11 | 0.0032679 | 0.0716407 | 0.046 | 0.964 |
| FA4_1.12 | 0.0908238 | 0.0935154 | 0.971 | 0.332 |
| FA4_2.12 | -0.0170404 | 0.0595895 | -0.286 | 0.775 |
| FA4_3.12 | 0.0684540 | 0.0796568 | 0.859 | 0.391 |
| FA4_4.12 | 0.0274375 | 0.0724595 | 0.379 | 0.705 |
| const | 0.0004067 | 0.0206003 | 0.020 | 0.984 |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.313 on 223 degrees of freedom
 Multiple R-Squared: 0.9033, Adjusted R-squared: 0.8998
 F-statistic: 260.3 on 8 and 223 DF, p-value: < 2.2e-16

Estimation resul for equation FA4_4:

```
=====
FA4_4 = FA4_1.11 + FA4_2.11 + FA4_3.11 + FA4_4.11 + FA4_1.12 + FA4_2.12 +
FA4_3.12 + FA4_4.12 + const
```

| | Estimate | Std. Error | t value | Pr(> t) |
|----------|-----------|------------|---------|--------------|
| FA4_1.11 | 0.189219 | 0.087057 | 2.174 | 0.0308 * |
| FA4_2.11 | -0.121155 | 0.056520 | -2.144 | 0.0331 * |
| FA4_3.11 | 0.100379 | 0.073746 | 1.361 | 0.1748 |
| FA4_4.11 | 1.314030 | 0.065742 | 19.988 | < 2e-16 *** |
| FA4_1.12 | -0.189856 | 0.085816 | -2.212 | 0.0280 * |
| FA4_2.12 | 0.108937 | 0.054683 | 1.992 | 0.0476 * |
| FA4_3.12 | -0.111736 | 0.073098 | -1.529 | 0.1278 |
| FA4_4.12 | -0.384480 | 0.066493 | -5.782 | 2.47e-08 *** |
| const | 0.008402 | 0.018904 | 0.444 | 0.6571 |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2872 on 223 degrees of freedom

Multiple R-Squared: 0.9186, Adjusted R-squared: 0.9157

F-statistic: 314.6 on 8 and 223 DF, p-value: < 2.2e-16

```
> library(KFAS)
> model<- SSModel(GDP~SSMregression(~ QFA4_1+QFA4_2+QFA4_3+QFA4_4,
Q=diag(NA,4)), H=NA)
> fit <- fitSSM(model, inits = c(0,0,0,0,0),method = "BFGS")
> out<- KFS(fit$model)
> print(out)
```

Smoothed values of states and standard errors at time n = 78:

| | Estimate | Std. Error |
|-------------|----------|------------|
| (Intercept) | 5.0574 | 0.2080 |
| QFA4_1 | -1.2427 | 0.9789 |
| QFA4_2 | 0.9886 | 1.0048 |
| QFA4_3 | -0.9591 | 0.6560 |
| QFA4_4 | 0.5378 | 0.9244 |

```
> model_DEF<- SSModel(DEF~ SSMregression(~ QFA4_1+QFA4_2+QFA4_3+QFA4_4,
Q=Qt), H=Ht)
```

```
> fit_DEF <- fitSSM(model_DEF, inits = c(0,0,0,0,0),method = "BFGS")
> out_DEF<- KFS(fit_DEF$model)
> print(out_DEF)
```

Smoothed values of states and standard errors at time n = 78:

| | Estimate | Std. Error |
|-------------|----------|------------|
| (Intercept) | 3.90991 | 0.10005 |
| QFA4_1 | 1.98670 | 0.49689 |
| QFA4_2 | 0.14295 | 0.11049 |
| QFA4_3 | 0.82816 | 0.10572 |
| QFA4_4 | -0.05745 | 0.10223 |

5 Comparison of forecasts and concluding remarks

Recognizing the need to utilize high-frequency indicators for more up-to-date forecasts, this paper has surveyed alternative modeling approaches to combining mixed frequency data for forecasting purposes. The models covered in this chapter include data-parsimonious (but more computer-demanding) models such as MF-DLFM as well as more data-intensive ones like CQM and MIDAS. In all these models, the fact that the data set is of mixed frequencies raises technical issues in the estimation and forecasting phases of the exercise. In the case of MF-DLFM, the additional feature of unobserved common factors introduces additional complications in implementing the estimation and simulation strategy based on the determination of a derived observable state-space formulation of the model.

The alternative models are estimated and constructed using Philippine data, to forecast GDP growth and inflation in the Philippines. For this numerical exercise, 10 monthly indicators are used for quarterly real GDP and 9 monthly indicators for the quarterly GDP deflator. The whole empirical analysis is implemented in R—starting from using R to access Philippine data from Philipiine and international data sources to analyzing the statistical properties of Philippine real GDP and GDP deflator and culminating in the estimation of the alternative forecasting models, where numerous variations of MIDAS are explored.

As the next step in this research agenda, it would be particularly important to compare the forecasting performance of the alternative procedures that are surveyed in this chapter. A more comprehensive study of this type will be presented in a future sequel to this chapter. For now, we proceed to summarize indicative comparison results recently reported in [Mariano and Ozmucur \(2018\)](#), where they compare one-period-ahead forecasts over the sample period for the following estimated models in their earlier paper: mixed-frequency dynamic latent factor model, MIDAS, Principal Components, Bridge Equations, and the benchmark AR and VAR models.

In terms of mean absolute errors and root mean square errors based on one-period ahead static forecasts, the results indicate that the MF-DLFM has the lowest mean absolute error—namely, 0.22% for real GDP growth rate and 0.28% for the GDP deflator growth rate. This is followed by MIDAS, with corresponding statistics of 0.45% and 0.37%. Principal Components and Bridge Equation follow these two models; and the benchmark AR and VAR models produce the biggest errors, almost four times as much as for MF-DLFM in forecasting real GDP growth.

[Mariano and Ozmucur \(2018\)](#) also applied the Diebold–Mariano test (see [Diebold and Mariano \(1995\)](#), [Mariano \(2002\)](#), and [Mariano and Preve \(2012\)](#)), to compare the statistical forecasting performance of MF-DLFM relative to the other models, taken one at a time. The test results are indicative of the significantly lower errors (at the 5% level of significance) for MF-DLFM—with

one exception, the MIDAS model for the GDP deflator growth rate. In this case, although errors are lower for MF-DLFM, the difference is not significant at the 5% level.

One more diagnostics in Mariano and Ozmucur (2018) assessed model performance in predicting turning points. All models do relatively well, if the prediction is for the level of GDP, real GDP or the GDP deflator. However, not all the estimated models fare well in predicting turning points in the growth rates. MF-DLFM correctly predicts 87% of turning points in real GDP while MIDAS predicts 74% of them; while the ratio is 79% for the Bridge and PCA models.

The results summarized here are at best indicative of the potentially superior performance of MF-DLFM for forecasting GDP growth, while for forecasting inflation, the performance of MF-DLFM is not significantly better than MIDAS. Thus more work is required for a more definitive conclusion on this issue—requiring further analysis and empirical applications, especially in expanding the performance analysis to cover the wider span of alternative forecasting models and variations of MF-DLFM and MIDAS surveyed in this chapter. Dynamic simulations for multiperiod forecasting also should be considered; as well as more refinements in the estimated models, especially the dynamic latent factor models, and extension of the analysis to other countries, especially in Southeast Asia.

Acknowledgments

The authors thank the University of Pennsylvania School of Arts and Sciences—for partial funding support; the United Nations DESA Expert Group Meeting on the World Economy (LINK Project) and the Sim Kee Boon Institute for Financial Economics at Singapore Management University—for providing seminar venues for presentation and discussion of this paper; and Michael Mariano—for his indispensable assistance in converting this chapter to LaTeX.

References

- Aruoba, S.B., Diebold, F.X., Scotti, C., 2009. Real-time measurement of business conditions. *J. Bus. Econ. Stat.* 27 (4), 417–427.
- Bennett, M.J., Hugen, D.L., 2016. *Financial Analytics With R, Building a Laptop Laboratory for Data Science*. UK, Cambridge.
- Coutino, A., 2005. A High-Frequency Model for Mexico. Project LINK web-site, <http://www.chass.utoronto.ca/LINK>.
- Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. *J. Bus. Econ. Stat.* 13, 253–265.
- Doz, C., Giannone, D., Reichlin, L., 2011. A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *J. Econ.* 164 (1), 188–205.
- Foroni, C., Marcellino, M., 2012. A Comparison of Mixed Frequency Approaches for Modelling Euro Area Macroeconomic Variables. *Economics Working Papers ECO 2012/07*, European University Institute.
- Foroni, C., Marcellino, M., 2013. A Survey of Econometric Methods for Mixed Frequency Data. *Economics Working Papers ECO 2013/02*, European University Institute.

- Ghysels, E., 2013. Matlab Toolbox for Mixed Sampling Frequency Data Analysis Using MIDAS Regression Models. Version 5. May, 2013.
- Ghysels, E., 2016a. MIDAS Matlab Toolbox. Version 2.1.
- Ghysels, E., 2016b. Macroeconomics and the reality of mixed-frequency data. *J. Econ.* 193, 294–314.
- Ghysels, E., Marcellino, M., 2018. Applied Economic Forecasting Using Time Series Methods. Oxford University Press.
- Ghysels, E., Santa-Clara, P., Valkanov, R., 2004. The MIDAS Touch: Mixed Data Sampling Regression Models. Working paper, Chapel Hill, NC.
- Ghysels, E., Sinko, A., Valkanov, R., 2007. MIDAS regressions: further results and new directions. *Econ. Rev.* 26 (1), 53–90.
- Ghysels, E., Kvedaras, V., Zemlys, V., 2016. Mixed frequency data sampling regression models: the R package midasr. *J. Stat. Softw.* 72 (4), 1–35, <https://www.jstatsoft.org/article/view/v072i04>.
- Gilbert, P., Meijer, E., 2015. Package ‘tsfa’—Time Series Factor Analysis. May 1, 2015, <https://cran.r-project.org/web/packages/tsfa/tsfa.pdf>.
- Harvey, A.C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.
- Heiss, F., 2016. Using R for Introductory Econometrics. Germany.
- Helske, J., 2018. Package ‘KFAS’—Kalman Filter and Smoother for Exponential Family State Space Models. September 19, 2018, <https://cran.r-project.org/web/packages/KFAS/KFAS.pdf>.
- Holmes, E., Ward, E., Scheuerell, M.D., Wills, K., 2018. Package ‘MARSS’—Multivariate Auto-regressive State-Space Modeling. November 2, 2018, <https://cran.r-project.org/web/packages/MARSS/MARSS.pdf>.
- Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting Principles and Practice, second ed. Otexts, Online, Open-Access Textbooks.
- Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F., R Core Team, Ihaka, R., Reid, D., Shaub, D., Tang, Y., Zhou, Z., 2019. Package ‘Forecast’—Forecasting Functions for Time Series and Linear Models. January 18, 2019, <https://cran.r-project.org/web/packages/forecast/forecast.pdf>.
- Inada, Y., 2005. A High-Frequency Model for Japan. Project LINK web-site, <http://www.chass.utoronto.ca/LINK>.
- Klein, L.R., Mak, W., 2005. University of Pennsylvania Current Quarter Model of the Chinese Economy. Forecast Summary. Project LINK web-site, <http://www.chass.utoronto.ca/LINK>.
- Klein, L.R., Ozmucur, S., 2001. The use of surveys in macroeconomic forecasting. In: Welfe, W. (Ed.), Macromodels 2001. University of Lodz, Poland.
- Klein, L.R., Ozmucur, S., 2002. Some possibilities for indicator analysis in economic forecasting. In: Project LINK Fall Meeting, University of Bologna, October 2002.
- Klein, L.R., Ozmucur, S., 2004. University of Pennsylvania Current Quarterly Model of the United States Economy Forecast Summary. Project LINK website, <http://www.chass.utoronto.ca/LINK>.
- Klein, L.R., Ozmucur, S., 2008. University of Pennsylvania high frequency macroeconomic modeling. In: Mariano, R.S., Tse, Y.-K. (Eds.), *Econometric Forecasting and High-Frequency Data Analysis*. Lecture Notes Series, vol. 13. World Scientific Publishers, Singapore, pp. 53–91. Institute for Mathematical Sciences, National University of Singapore. 2008. Earlier version presented at the High Frequency Modeling Conference, Singapore Management University (SMU), Singapore, May 7–8, 2004.
- Klein, L.R., Park, J.Y., 1993. Economic forecasting at high-frequency intervals. *J. Forecast.* 12, 301–319.

- Klein, L.R., Park, J.Y., 1995. The University of Pennsylvania model for high-frequency economic Forecasting. In: Economic and Financial Modelling, vol. 2, pp. 95–146. Autumn 1995.
- Klein, L.R., Sojo, E., 1987. Combinations of high and low frequency data in macroeconomic models. In: Paper Presented at the Session “Can Economic Forecasting Be Improved?” American Economic Association. Chicago, Illinois. December.
- Klein, L.R., Sojo, E., 1989. Combinations of high and low frequency data in macroeconomic models. In: Klein, L.R., Marquez, J. (Eds.), Economics in Theory and Practice: An Eclectic Approach. Kluwer Academic Publishers, pp. 3–16.
- Klein, L.R., Eskin, V., Roudoi, A., 2003. Empirical regularities in the Russian economy. In: Project LINK Spring Meeting. United Nations, New York, April 23–25, 2003.
- Klein, L.R., Eskin, V., Roudoi, A., 2005. University of Pennsylvania and Global Insight Current Quarter Model of the Russian Economy. Forecast Summary. Project LINK web-site, <http://www.chass.utoronto.ca/LINK>.
- Kvedaras, V., Zemlyns, V., 2016. Package ‘midasr’—Mixed Data Sampling Regression, August 16, 2016. CRAN (Comprehensive R Archive Network), <https://cran.r-project.org/web/packages/midasr/midasr.pdf>.
- Liu, H., Hall, S.G., 2001. Creating high frequency national accounts with state-space modelling: a Monte Carlo experiment. *J. Forecast.* 20, 441–449.
- Luethi, D., Erb, P., Otziger, S., 2018. Package ‘FKF’—Fast Kalman Filter. July 20, 2018, <https://cran.r-project.org/web/packages/FKF/FKF.pdf>.
- Mariano, R.S., 2002. Testing forecast accuracy. In: Clements, M.P., Hendry, D.F. (Eds.), A Companion to Economic Forecasting. Blackwell, Oxford.
- Mariano, R.S., Murasawa, Y., 2003. A new coincident index of business cycles based on monthly and quarterly series. *J. Appl. Econom.* 18 (4), 427–443.
- Mariano, R.S., Murasawa, Y., 2010. A coincident index, common factors, and monthly real GDP. *Oxf. Bull. Econ. Stat.* 72 (1), 27–46.
- Mariano, R.S., Ozmucur, S., 2018. High-mixed-frequency forecasting models for GDP and inflation. Ch. 1, In: Pauly, P. (Ed.), Global Economic Modeling—A Volume in Honor of Lawrence Klein. World Scientific Publishing Co. Pt. Ltd, pp. 2–29.
- Mariano, R.S., Preve, D., 2012. Statistical tests for multiple forecast comparison. *J. Econ.* 169 (1), 123–130.
- Mariano, R.S., Tse, Y.-K. (Eds.), 2008. Econometric Forecasting and High-Frequency Data Analysis, In: Lecture Notes Series, vol. 13. World Scientific Publishers, Singapore. Institute for Mathematical Sciences, National University of Singapore.
- Ozmucur, S., 2009. Current quarter model for Turkey. In: Klein, L.R. (Ed.), The Making of National Economic Forecasts. Edward Elgar Publishing Ltd., Cheltenham, UK/Northampton, MA, USA, pp. 245–264, 2009, Chapter 9.
- Pauly, P. (Ed.), 2018. Global Economic Modeling—A Volume in Honor of Lawrence Klein. World Scientific Publishing Co. Pt. Ltd.
- Petris, G., 2018. Package ‘dlm’—Bayesian and Likelihood Analysis of Dynamic Linear Models. June 13, 2018, <https://cran.r-project.org/web/packages/dlm/dlm.pdf>.
- Pfaff, B., Stigler, M., 2018. Package ‘vars’—VAR Modelling. August 6, 2018, <https://cran.r-project.org/web/packages/vars/vars.pdf>.
- R Core Team, 2018. The R Stats Package—R Statistical Functions. CRAN (Comprehensive R Archive Network), <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html>.
- R Development Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

- Reinhart, A., 2017. Package ‘pdfetch’. October 15, 2017, <https://cran.r-project.org/web/packages/pdfetch/pdfetch.pdf>.
- Revelle, W., 2019. Package ‘psych’—Procedures for Psychological, Psychometric, and Personality Research. January 13, 2019. <https://cran.r-project.org/web/packages/psych/psych.pdf>.
- Sargent, T., Sims, C., 1977. Business cycle modeling without pretending to have too much a priori economic theory. In: New Methods in Business Cycle Research. Federal Reserve Bank of Minneapolis.
- Shumway, R.H., Stoffer, D.S., 2017. Time Series Analysis and Its Applications, With R Examples, fourth ed. Springer International Publishing, Switzerland.
- Stock, J.H., Watson, M.W., 1989. New indexes of coincident and leading economic indicators. In: Blanchard, O.J., Fischer, S. (Eds.), NBER Macroeconomics Annual, vol. 4. MIT Press, Cambridge, Massachusetts, pp. 351–409.
- Stoffer, D., 2017. Package ‘astsa’—Applied Statistical Time Series Analysis. December 15, 2017, <https://cran.r-project.org/web/packages/astsa/astsa.pdf>.
- Vinod, H.D., 2011. Hands-on Intermediate Econometrics Using R, Templates for Extending Dozens of Practical Examples. World Scientific Publishing Co., Singapore
- Wickham, H., Bryan, J., Kalicinski, M., Valery, K., Leitienne, C., Colbert, B., Hoerl, D., Miller, E., 2018. Package ‘Readxl’—Read Excel Files. December 20, 2018, <https://cran.r-project.org/web/packages/readxl/readxl.pdf>.

Further reading

- Philippine Statistics Authority National Statistical Coordination Board, 2014. Leading Economic Indicators. First Quarter 2014, http://www.nscb.gov.ph/lei/publication/PSA-NSCB_LEI%20Q12014.pdf.
- Stock, J.H., Watson, M.W., 2005. Implications of Dynamic Factor Models for VAR Analysis. National Bureau of Economic Research Working Paper 11467. June 2005.

Chapter 7

Nonlinear time series in R: Threshold cointegration with tsDyn

Matthieu Stigler*

*Department of Agricultural and Resource Economics, University of California, Davis, CA,
United States*

*Corresponding author: e-mail: matthieu.stigler@gmail.com

Abstract

The notion of cointegration describes the case when two or more variables are each nonstationary, yet there exists a combination of these variables which is stationary. This statistical definition leads to a rich economic interpretation, where the variables can be thought of as sharing a stable relationship, and deviations from a long-run equilibrium are corrected. Implicit in the definition, however, is the requirement that any small deviation from the long-run equilibrium needs to be corrected instantaneously and symmetrically.

Threshold cointegration relaxes the linear and instantaneous adjustment assumption by allowing the adjustment to occur only after the deviations exceed a critical threshold. Likewise, it can accommodate asymmetric adjustment, where positive or negative deviations are not necessarily corrected in the same way. This flexible framework can be used to model economic phenomena such as transaction costs, stickiness of prices, or asymmetry in agents' reactions.

In this chapter, I survey the concept of threshold cointegration, and show how to use this model within R with package **tsDyn**. In Section 1, I review briefly the concept of stationarity and cointegration, then explain the concept of threshold cointegration. In Section 2, I discuss in detail the econometrics of threshold cointegration, presenting the main tests and estimators. I describe then the package **tsDyn** in Section 3, and show how to use it with an empirical application on the term structure of interest rates in Section 4.

Keywords: Nonlinear time series, Nonstationarity, Cointegration, Error correction, Threshold cointegration, Regime-switching, Term structure, Self-exciting threshold autoregressive model

1 Introduction: Linear and threshold cointegration

An important stylized fact in the analysis of time series is that many economic variables are nonstationary, following a random-walk type behavior. [Samuelson \(1965\)](#) argued that prices should follow a random walk. [Nelson and Plosser \(1982\)](#) investigated a set of 14 macroeconomic series, finding that all but one of them were *difference-stationary*. Difference-stationary refers to the case where a variable y_t is nonstationary, but whose difference Δy_t is stationary and is also said to be *integrated of order 1* (or sometimes just integrated, abbreviated I(1)), or to have a unit root. The random walk $y_t = \rho y_{t-1} + \epsilon_t$ with $\rho = 1$ and ϵ_t i.i.d. is a leading example of an I(1) process.^a

At the econometric level, I(1) variables have different properties than the usual stationary variables, and require a different set of tools for inference. In particular, regression among integrated series leads to a spurious regression, i.e., inflation of regression indicators (t -tests, R^2) which lead to the false conclusion of statistical dependence between variables ([Granger and Newbold, 1974](#); [Phillips, 1986](#)). An obvious remedy in this case is to use differenced series, for which usual asymptotics apply. This approach became the standard in the vector autoregression (VAR) framework popularized by [Sims \(1980\)](#).

The concept of cointegration, introduced by [Engle and Granger \(1987\)](#), denotes an interesting case where **variables are nonstationary, yet there exists a linear combination that is stationary**. This can be interpreted economically as **the presence of a long-run equilibrium**. Cointegration gained popularity with the **Granger representation theorem**, which states that **cointegrated variables have a vector error correction model (VECM) representation**. A VECM is a VAR in differences including an additional variable representing the deviations from the long-run equilibrium. To illustrate this, consider the case of two I(1) variables, Y_t and X_t , and the **cointegrating value β** , such that $ECT_t \equiv Y_t - \beta X_t$ is stationary. The VECM representation is shown in Eq. (1) with a single lag:

$$\begin{bmatrix} \Delta X_t \\ \Delta Y_t \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} ECT_{t-1} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} \Delta X_{t-1} \\ \Delta Y_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t^X \\ \varepsilon_t^Y \end{bmatrix}. \quad (1)$$

This VECM representation is particularly interesting as it allows us to estimate **how the variables adjust to deviations from the long-run equilibrium** (**the α parameters**), **to test for Granger-causality**, and to determine the impacts of shocks to the variables using **impulse response functions**. The case $k > 2$ follows along almost the same lines, with the important difference that there can be more than one cointegrating relationship, and hence more than one ECT term.

Since [Engle and Granger \(1987\)](#), cointegration has become a popular tool in the analysis of integrated time series, and has been applied in a wide variety of settings. Part of the popularity of the concept comes from the fact that it has

^aIndeed this process is stationary when taking its difference: $\Delta y_t \equiv y_t - y_{t-1} = \epsilon_t$, which is stationary by definition of ϵ_t .

a sound interpretation in terms of variables following a long-term common relationship, and adjusting to deviations from that relationship. [Balke and Fomby \(1997\)](#) note, however, that implicit in this concept is the assumption that the adjustment of the deviations toward the long-run equilibrium is made **instantaneously, and symmetrically**. There are nevertheless various arguments in economic theory challenging this assumption of instantaneous adjustment. As a leading example, the presence of transaction costs implies that adjustment will occur only once deviations are higher than the transactions costs. Financial theory predicts that even in highly liquid markets a band of no arbitrage may exist where deviations are too small for the arbitrage to be profitable, if there are transaction costs. In the domain of macroeconomics, policies are often organized around targets, where intervention is activated only once the deviations from the target are significant, the most famous example being monetary policy management during the Bretton Woods agreement where central banks pegged their exchange rate to the dollar and allowed a $\pm 1\%$ band. Likewise, the assumption of symmetric adjustment implicit in the linear cointegration model is challenged by various economic arguments. [Levy et al. \(1997\)](#) and [Dutta et al. \(1999\)](#) argue that the presence of menu costs induces asymmetric price adjustments, while [Damania and Yang \(1998\)](#) and [Ward \(1982\)](#) predict that market power leads to an asymmetric price transmission.

To take these criticisms of linear cointegration into account, [Balke and Fomby \(1997\)](#) introduced the concept of **threshold cointegration**. In their framework, the adjustment does not need to occur instantaneously but only once the deviations exceed some critical threshold, allowing thus the presence of an inaction or no-arbitrage band. They base their adjustment process on the ***self-exciting threshold autoregressive model (SETAR)*** introduced by [Tong \(1978\)](#) and discussed extensively in [Tong \(1990, 2011, 2015\)](#). In the SETAR model, the **autoregressive coefficients take different values depending on whether the previous value is above or under a certain threshold value**. Hence, the linear AR(1) process:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \quad (2)$$

and is modified into the SETAR(2) process with two thresholds, θ_L and θ_H :

$$\varepsilon_t = \begin{cases} \rho_L \varepsilon_{t-1} + u_t & \text{if } \varepsilon_{t-1} \leq \theta_L \\ \rho_M \varepsilon_{t-1} + u_t & \text{if } \theta_L < \varepsilon_{t-1} \leq \theta_H \\ \rho_H \varepsilon_{t-1} + u_t & \text{if } \theta_H < \varepsilon_{t-1} \end{cases} \quad (3)$$

This is actually a **piecewise linear model** where three different AR(1) processes are estimated depending on the state of the variable ε at time $t-1$, i.e., whether we have the low regime $\varepsilon_{t-1} \leq \theta_L$, the middle regime $\varepsilon_{t-1} \in]\theta_L, \theta_H]$, or the high regime $\theta_H < \varepsilon_{t-1}$. This SETAR model clearly nests the AR one (if $\rho_H = \rho_M = \rho_L$), and can easily be extended to allow for lags and intercepts. Note that if one were to use time t as threshold variable instead of ε_{t-1} , one would obtain the usual structural break model.

The innovation of Balke and Fomby was to apply this SETAR model to the cointegration case, where now ϵ_t represents the error correction term, i.e., the deviations from the long-run relationship $\epsilon_t \equiv y_t - \beta x_t$. This allows to relax linear cointegration in two ways. First of all, asymmetric adjustment can be modeled with a SETAR(1) model with one threshold $\theta=0$, and $\rho_L \neq \rho_H$. Second, an interesting feature of the SETAR model is that it can be globally stationary despite being nonstationary in some regimes. For example, if the inner regime contains a unit root ($\rho_M=1$), yet the outer regimes are stationary ($\rho_L, \rho_H < 1$), the model is globally stationary.^b This corresponds nicely to the economic idea of a *band of inaction*, where small deviations fluctuate randomly, yet larger ones lead to mean-reversion. Balke and Fomby considered even more general cases of SETAR models, where adjustment is not toward a point but a given band, which they called BAND-TAR. This model is, however, more complicated, and received less treatment in the literature, and will not be further discussed here.

While the work of Balke and Fomby (1997) focused on the long-run cointegrating relationship (which I will denote as *residual-based* approach), extension to a threshold VECM (TVECM) has been made by Hansen and Seo (2002) and Seo (2006) among others.^c The VECM model shown in Eq. (1) can be generalized to the TVECM, where the speed of adjustment parameters α change depending on the values of the error correction term (ECT):

$$\begin{bmatrix} \Delta X_t \\ \Delta Y_t \end{bmatrix} = c + \alpha^L ECT_{t-1}^L + \alpha^M ECT_{t-1}^M + \alpha^H ECT_{t-1}^H + B \begin{bmatrix} \Delta X_{t-1} \\ \Delta Y_{t-1} \end{bmatrix} + \epsilon_t \quad (4)$$

where the error-correction term ECT is split into three regimes, lower (L), middle (M) and high (H) depending on whether it is below, between or above the two thresholds θ_L and θ_H , as in Eq. (3). As in the TAR model, the TVECM nests the VECM model if $\alpha^L = \alpha^M = \alpha^H$, and can be generalized by adding more lags, and allowing all parameters to change depending on the ECT value.

Note that *threshold cointegration* here refers either to deviations from the (linear) cointegration relationship following a SETAR model, or the parameters in a VECM having different values depending on the deviations from the linear cointegration relationship. Other specifications, where the threshold effect is not the ECT term, have been considered by Gonzalo and Pitarakis (2006a) and Krishnakumar and Neto (2015). Gonzalo and Pitarakis (2006b) consider yet another case, where the threshold effect arises through different parameters β in the cointegrating relationship, while the deviations from this relationship follow a usual linear AR model.

^bSee Chan et al. (1985) for more general conditions.

^cGranger and Lee (1989) are an early example of testing for asymmetric adjustment between negative and positive deviations.

Since the seminal work of Balke and Fomby, threshold cointegration has become widely applied in various contexts. The law of one price (LOP) and the associated purchasing power parity (PPP) hypothesis represent the fields where the most studies have been conducted (see for a review on LOP Lo and Zivot, 2001; for the PPP, confer Bec et al., 2004; Gouveia and Rodrigues, 2004; and Heimonen, 2006). The threshold cointegration frame-work has been intensively used to study price transmission of agricultural products or other commodities (principally oil); see, for example, Sephton (2003), Cramon-Taubadel (1998), Meyer (2004), and Sergio et al. (2017) for agricultural products; Ghassan and Banerjee (2015), Hammoudeh et al. (2008), and Zhu et al. (2011) for oil markets. Other domains include the term structure of interest rates (Krishnakumar and Neto, 2008), the Fisher effect (Million, 2004), stock market integration (Jawadi et al., 2009), exchange rate pass-through (Al-Abri and Goodwin, 2009). To my knowledge, the use of thresh-old cointegration remains within the economics literature and no work has been done in other domains.

2 Estimation, testing, and interpretation

Although the threshold cointegration model delivers interesting insights, estimation and testing are fairly complicated. It combines cointegration and threshold models that each present several difficulties on their own. In Section 2.1, I highlight the main econometric complications arising with cointegration, then discuss the threshold model in Eq. (5), and finally show how these combine in the threshold cointegration case in Section 2.3.

2.1 Estimation and testing for cointegration

Two main approaches are used to estimate and test for cointegration. The first, advocated in the seminal paper of Engle and Granger (1987), is a two-step approach, which I will refer to as the residual-based approach. In the first step, a simple regression using the long-run relationship is estimated with OLS. Assuming we have just two nonstationary variables, y_t and x_t , the long-run relationship is:

$$y_t = \beta x_t + \varepsilon_t$$

this gives the cointegrating values $\hat{\beta}$. The ECT term is constructed as the residual from this regression, $ECT_t = \hat{\varepsilon}_t$ (hence the alternative name, residual-based approach). To test for cointegration, a unit root test is applied to the estimated ECT term. If the unit root test is rejected, one concludes that there is cointegration among the variables, and plugs the ECT into the VECM as the second step. This two-step approach is justified asymptotically by the fact that the estimator $\hat{\beta}$ in the first step is super-consistent—it converges at the rate of n instead of the usual rate \sqrt{n} (see Stock, 1987). This means that

standard inference can be used for the parameters of the VECM, proceeding as if the cointegration parameters β were known.

Testing for cointegration in the residual-based approach amounts to testing for stationarity of the $ECT_t = \hat{\epsilon}_t$ term. This is done traditionally using a unit root test approach, which has as null hypothesis the unit root, $H_0: \rho = 1$ vs $H_A: \rho < 1$ (where ρ is the AR(1) parameter in $\hat{\epsilon}_t = \rho \hat{\epsilon}_{t-1} + u_t$). If the cointegrating vector is known a priori, the usual critical values from the unit root test can be used. In the general case where the cointegrating values are not known and need to be estimated in the first step, critical values need to be adjusted for the uncertainty in the first step, see [Phillips and Ouliaris \(1990\)](#).^d

A major drawback of the two-step approach is that it allows estimation and testing for only one single cointegrating relationship. However, if there are more than two variables, there can be more than one cointegrating relationship.^e

[Johansen \(1988, 1996\)](#) developed a maximum-likelihood procedure to estimate and test for multiple cointegrating relationships in the *system-based* VECM.^f

2.2 Estimation and testing for threshold models

Turning now to estimation and testing for threshold models, let us start with a simple SETAR(1) model with one threshold, an intercept, and a single lag in each regime:

$$y_t = (\mu_L + \rho_L y_{t-1}) \mathbb{1}(y_{t-1} \leq \gamma) + (\mu_H + \rho_H y_{t-1}) \mathbb{1}(y_{t-1} > \gamma) + \varepsilon_t. \quad (5)$$

The $\mathbb{1}(\cdot)$ is the indicator function taking a value of 1 if the event (\cdot) is true, and 0 otherwise. Estimation of the model usually uses least-squares, minimizing the objective function:

$$SSR(\theta, \gamma) \equiv \sum_t \varepsilon_t(\theta, \gamma)^2 \quad (6)$$

where $\varepsilon_t(\theta, \gamma)$ denotes the residuals from model Eq. (5) with threshold parameter γ and slope parameter $\theta \equiv (\mu_L, \mu_H, \rho_L, \rho_H)$. The discontinuity of the indicator function precludes deriving an analytical form for Eq. (6), or using standard numerical optimization algorithms. A solution can be found noting that estimation of the slope parameters θ is easy in case of a given threshold: it is simply OLS, which we denote as $\hat{\theta}(\gamma)$. This suggests that by precomputing $\hat{\theta}(\gamma)$ for a given γ , we can reduce the multi-dimensional problem (θ, γ) to a one-dimensional problem that depends only on γ :

$$\hat{\gamma} = \arg \min_{\gamma} SSR(\gamma) = SSR(\hat{\theta}(\gamma), \gamma). \quad (7)$$

^dAvailable in R in package **urca** ([Pfaff, 2008a](#)).

^eIn fact, the number of maximum possible cointegrating relationships is K (number of variables) – 1.

^fThis is available in R with the `ca.jo()` function from package **vars** ([Pfaff, 2008b](#)).

This procedure is usually referred to as concentrated least squares, or also profiling. Minimization of Eq. (7) is done through a grid search over the observed values of the threshold variable.^g In practice, the values of the threshold variable are sorted and a certain percentage of the first and last values is excluded to ensure a minimal number of observations in each regime. The SSR is then evaluated for each candidate value and the one that minimizes the SSR is taken as the estimator. Finally, once the threshold parameter $\hat{\gamma}$ has been obtained, the corresponding slope parameters $\hat{\theta}(\hat{\gamma})$ are estimated by OLS.

The procedure for two thresholds can be conducted in the same way, and searching on all combinations of γ_L, γ_H to minimize $SSR(\gamma_L, \gamma_H, \hat{\theta}(\gamma_L, \gamma_H))$. This is, however, a T^2 dimensional search and may rapidly become cumbersome. A computational shortcut was suggested in [Balke and Fomby \(1997\)](#). The idea is to estimate the threshold in a sequential way: the search is done first in a model with only one threshold. The second threshold is then estimated taking the first as given. A few iterations can be conducted, reestimating the first threshold conditional on the second one and vice versa. [Gonzalo and Pitarakis \(2002\)](#) showed that this algorithm is consistent as the estimator in the first step of the misspecified model is nevertheless consistent for one of the thresholds. This is a substantial shortcut as it reduces the number of computations from T^2 to $2 \times T$, or $(2+k) \times T$ if k iterations are done, practice showing that after two or three iterations a maximum is reached.

Properties of the concentrated LS estimator described above were obtained by [Chan \(1993\)](#). He established that the estimator of the threshold, $\hat{\gamma}$, is super-consistent (i.e., converges at rate T instead of the usual \sqrt{T}). The superconsistency of $\hat{\gamma}$ implies that for the slope coefficients $\hat{\beta}$, one can proceed as if the true γ_0 value was known, and hence the asymptotic distribution of $\hat{\beta}(\hat{\gamma})$ is independent of $\hat{\gamma}$ and is the same as that of $\hat{\beta}(\gamma_0)$, the standard normal distribution in the case of OLS. Chan shows also that the distribution of $\hat{\gamma}$ is a *compound Poisson process*. This is a complicated distribution that depends on a host of nuisance parameters, so that it cannot be used easily to derive confidence intervals for $\hat{\gamma}$. Several papers have offered alternative solutions. [Hansen \(2000\)](#) uses an a *vanishing-threshold* asymptotic framework, under which he obtains confidence intervals for the threshold parameter. [Gonzalo and Wolf \(2005\)](#) use subsampling procedures. [Seo and Linton \(2007\)](#) smooth the objective function, slowing the convergence rate of the estimator, but yielding a normal distribution for $\hat{\gamma}$. [Li and Ling \(2012\)](#) propose algorithms to simulate the compound Poisson process.

^gNote that one needs indeed only to look at observed values of the threshold variable as candidates: for any two sorted values $y_{(i)} < y_{(i+1)}$, any value γ in the interval $y_{(i)} < \gamma < y_{(i+1)}$, $\mathbb{1}(y_t \leq \gamma) = \mathbb{1}(y_t \leq y_{(i+1)})$.

An important preliminary step in the analysis of threshold models is testing for the presence of a threshold effect; i.e., testing $\theta^L \neq \theta^H$. In case of a known threshold parameter, a standard likelihood-ratio test for equality of coefficients can be used (Chan and Tong, 1990), similar to using a Chow test for a structural break with known change date. But when the threshold is unknown—which is typically the case in practice—one faces a nonstandard testing scenario: the alternative hypothesis depends on $\hat{\gamma}$, while the null does not. This is a case of testing with a *parameter not identified under the null*, a problem studied early on by Davies (1977, 1987).^h

Solutions for that problem (Andrews and Ploberger, 1994) involve computing the test statistic over a range of possible threshold values and then aggregating those results with a function like the mean or the supremum. The asymptotic distribution of the resulting statistic is unfortunately quite complicated in the threshold case, so that bootstrap methods need to be used (see Hansen, 1996, 1999). The use of bootstrap methods makes the test quite computationally intense: one needs to estimate the test statistic each time over a large range of values, and repeat this over a large number of bootstrap replications. Noting the equivalence between model selection and testing, Gonzalo and Pitarkis (2002) suggest using information criteria such as the AIC, BIC, or modifications thereof to decide on the number of regimes. They show that this model selection approach consistently determines the number of regimes. This substantially reduces the number of computations, yet leaves open the question of which criterion to use in practice.

An alternative approach is to turn the problem into a test for structural change. If observations are sorted according to the threshold values, instead of being sorted according to time, estimation and testing for a threshold in the original regression is equivalent to estimation and testing for a break in the *arranged* regression. This allows the use of structural break tests, which have received more attention in the literature. Structural break tests, despite sharing the same issue of a parameter not identified under the null as in the threshold case, have the advantage that the resulting test statistics are free of nuisance parameters (Andrews, 1993). This was used in early tests for threshold nonlinearity, see Petrucci and Davies (1986) and Tsay (1989). Furthermore, this approach is also useful for estimation of the thresholds, as efficient algorithms have been developed in the structural change literature (see, for example, Bai and Perron, 2003 Killick et al., 2012). While promising, the equivalence between arranged threshold models and structural change models unfortunately only holds when the observed values are all distinct: if any value repeats, there is no natural ordering given the presence of ties (Hansen, 2000, p. 578).

^hThis problem arises in a wide variety of settings. Think, for example, of a test for omitted nonlinearity: $y_i = \beta x_i + \delta(x_i)^\alpha + \varepsilon_i$, where α is a power of x . If one sets a priori $\alpha=2$, one can use a usual t -test for $\delta=0$. However, if one wants to test for a general α , α is not defined in the case $\delta=0$.

2.3 Estimation and testing for threshold cointegration models

As discussed in Sections 2.1 and 2.2, the econometrics of cointegration and of threshold models are quite involved each on their own. It shall thus come as no surprise that the threshold cointegration model faces a host of challenges for estimation and testing.

As noted early on by [Balke and Fomby \(1997\)](#), testing for threshold cointegration involves testing for both the presence of cointegration and threshold nonlinearity. Hence there are **four cases**:

1. **Cointegration and threshold effects.** This is the **threshold cointegration case**.
2. **Cointegration and no threshold effects.** This is the **linear cointegration case**
3. **No cointegration and no threshold effects.** This is the **no cointegration case**.
4. Threshold effects and no cointegration.

The last case, threshold noncointegration, has received less attention in the literature, so will not be discussed further. We have then two alternatives to **threshold cointegration: no cointegration** (case 3), or **linear cointegration** (case 2). This suggest two approaches to test for threshold cointegration:

- **Two-step approach:**
 1. **Test for no cointegration vs cointegration**
 2. **If cointegration is found, test cointegration vs threshold cointegration**
- **Direct approach:** test no cointegration vs threshold cointegration.

In their seminal paper, Balke and Fomby explored the two-step approach. It has the advantage that for the second step, stationarity can be assumed, which makes the test simpler. It has the drawback, however, that it suffers from low power: even if there is threshold cointegration, the first step might not point toward cointegration, in which case we would conclude that there is no threshold cointegration. The direct approach does not suffer from this drawback, yet the testing procedure becomes more complicated, as under the null hypothesis, the process is a unit root test.

To illustrate the two approaches, consider the case of a residual-based test, assuming a known cointegrating vector β , $y_t = \beta x_t + \epsilon_t$. The AR(1) process based on the residuals is $\epsilon_t = \rho \epsilon_{t-1} + u_t$, whereas a SETAR model with two regimes is: $\epsilon_t = \rho^L \epsilon_{t-1} \mathbb{1}(\epsilon_{t-1} \leq \gamma) + \rho^H \epsilon_{t-1} \mathbb{1}(\epsilon_{t-1} > \gamma) + u_t$.

- Two-step approach:
 1. **RW vs AR:** Test first for (no) cointegration: $H_0 : \rho = 1$ vs $H_A : \rho < 1$
 2. **AR vs SETAR:** If rejected, test for threshold effects: $H_0 : \rho^L = \rho^H | \rho < 1$ vs $H_A : \rho^L \neq \rho^H$
- The direct approach:
 1. **RW vs SETAR:** $H_0 : \rho^L = \rho^H = 1$ vs $H_A : \rho^L < 1$ and $\rho^H < 1$

TABLE 1 Threshold tests: different approaches

| Approach | | Hypothesis | | | Known β vector? | |
|----------|----------|------------|----|-------|------------------------------|------------------------|
| | | H_0 | | H_A | Yes | No |
| Two-step | Residual | (if) AR | vs | SETAR | Hansen (99) | |
| | System | (if) VECM | vs | TVECM | | Hansen, Seo (02) |
| Direct | Residual | RW | vs | SETAR | Bec et al. (01), Seo (08) | Enders, Siklos (01) |
| | | | | | Kapetanios, Shin (06) | |
| | System | VAR | vs | TVECM | Seo (06) | |

Note: RW stands for random-walk.

The same procedure could be repeated with the system-based approach, where one will test for VAR vs VECM, then VECM vs TVECM (two-step approach) or VAR vs TVECM (direct approach). Unfortunately, some of these approaches are available only in the case of a known cointegrating vector. [Table 1](#) summarizes the different cases, and the main available tests implemented in [tsDyn](#).

2.3.1 Two-step approach

[Balke and Fomby \(1997\)](#) follow the two-step approach, using residual-based tests. They test first for cointegration by estimating the long-run relationship with OLS, and then apply unit root tests on the residuals. Whenever they reject the unit root tests they then test for threshold effects, using a stationary model as the null hypothesis. The authors also try the two-step approach using VECM-based Johansen cointegration tests. Results of their Monte-Carlo simulations show that overall, standard linear cointegration tests have more power when the alternative is a stationary threshold autoregression.

A test using the system-based two-step approach is suggested by [Hansen and Seo \(2002\)](#). Given a bivariate cointegrated VECM, the authors test for threshold effects in the error correction term as well as lag parameters. Denoting the parameters in the low and high regimes A_L and A_H , respectively, they test $H_0: A_H = A_L$ against $H_A: A_H \neq A_L$ using a sup-LM test. The distribution of the sup-LM test is found to be the same as in the univariate case of [Hansen \(1996\)](#). This distribution cannot be tabulated due to the presence of nuisance parameters and hence the authors suggest two bootstrap approaches, with either a fixed-regressor or a residual bootstrap.

2.3.2 Direct test: No cointegration vs threshold cointegration

As argued before, an alternative to the two-step approach is to use tests with no cointegration as the null hypothesis. This, however, raises the difficulty that under the null hypothesis, the residuals (or the error correction term under the system-based approach) are now nonstationary.

For residual-based direct tests, an important distinction is to be made between test assuming a known cointegrating vectors or not. In the case of a known cointegrating vector, the problem reduces to testing for a random-walk vs a stationary SETAR(1) to establish threshold cointegration. Multiple tests have been developed in this context; see, for example, [Bec et al. \(2008\)](#), [Kapetanios and Shin \(2006\)](#), [Seo \(2008\)](#), or [Maki \(2009\)](#) for a review. In the general case where the cointegrating vector is not known, [Enders and Siklos \(2001\)](#) derive a test and provide critical values for it. These critical values were, however, derived from a specific Monte-Carlo experiment with no support from formal theory, so it is not clear how their specific Monte-Carlo setup generalizes.

In the class of system-based direct tests, only a test with known cointegrating value is available, to my knowledge. This is the test of [Seo \(2006\)](#). [Seo \(2006\)](#) notes that if the cointegrating vector is known, a linear cointegration test can be done by testing that the coefficients for the error correction term in the VECM (the α parameters in Eq. 1) are different from 0, i.e., test the null of $H_0: \alpha = 0$ (see [Horvath and Watson, 1995](#)). Extension to the threshold case implies testing $H_0: \alpha^L = \alpha^H = 0$. [Seo \(2006\)](#) proposes a sup-Wald test for this hypothesis, providing critical values as well as a bootstrap procedure. This procedure unfortunately works with a known cointegrating vector. One should note, however, that, by using a slight modification of the model,ⁱ [Li and Lee \(2010\)](#) and [Li \(2017\)](#) derive system-based tests that are valid when the cointegrating vector is estimated.

2.3.3 A small remark on testing for threshold cointegration

It is useful to remember that stationarity of a SETAR process holds under various sets of conditions. All the tests discussed here assume that the outer regimes of the SETAR are stationary. This is, however, only a sufficient condition, and not a necessary one. Indeed, depending on the values of the intercepts, a SETAR process could have unit roots in the outer regimes and yet be stationary. This implies a complicated set of conditions to test; to my knowledge, there is no available test for this. Once may hence conjecture that some of the series described as nonstationary in the literature may well be SETAR-stationary.

ⁱThe model assuming that the regime-switching is based on a percentile of the threshold variable, instead of a fixed value. Furthermore, they investigate the case of MTAR adjustment, not the TAR adjustment that has been discussed here.

2.4 Estimation of a threshold estimated model

To estimate the parameters of a threshold cointegration model, the literature has used both two-step and joint estimation procedures. Balke and Fomby (1997) and other early work used a linear model to estimate the cointegration parameter, then estimated the threshold parameter using a grid search on the residuals, as outlined in Section 2.2. Hansen and Seo (2002) suggested a joint estimation procedure, where the cointegration and threshold parameters are estimated together using a two-dimensional grid. While the space for the threshold parameter is restricted to existing values of the threshold, the parameter space for β is unrestricted. Hansen and Seo suggested setting a grid around the confidence interval of the β obtained with the linear estimator. The drawback of this joint estimation approach is that the grid search becomes quickly infeasible with more than two variables, so that estimation has been largely restricted to bivariate VECMs (although procedures based on genetic algorithms have been suggested, see El-Shagi, 2011). Seo (2011) formally investigated estimation of threshold cointegration. While his findings cast doubts on the validity of the two-step approach, Seo proves the consistency of the joint estimation procedure. Interestingly, superconvergence of the cointegration parameter not only holds, but is even stronger: the $\hat{\beta}$ estimator converges at the super fast rate of $T^{3/2}$. The $\hat{\gamma}$ estimator converges again at the fast rate of T , yet has a complicated distribution including nuisance parameters. Seo shows that the smoothed least-squares method from Seo and Linton (2007) can be applied in the threshold cointegration case, leading to a normal distribution for the resulting smoothed LS estimator $\hat{\gamma}$.

2.5 (Generalized) impulse response functions

As soon as they involve more than one lag, univariate and multivariate time series models are difficult to interpret. One way to investigate the dynamics of a model is to look at the so-called impulse response function (IRF). The function $IRF(h, \delta) = \frac{\partial y_{t+h}}{\partial \epsilon_t}$ shows how a shock $\epsilon_t = \delta$ at time t impacts a system at time $t+h$, assuming no further shocks, $\epsilon_{t+h} = 0 \quad \forall h$. It can be also thought as comparing two concurrent realizations, one with the shock, one without it:

$$IRF(h, \delta, \omega_{t-1}) = E[y_{t+h} | \epsilon_t = \delta, \epsilon_{t+h} = 0, \omega_{t-1}] - E[y_{t+h} | \epsilon_t = 0, \epsilon_{t+h} = 0, \omega_{t-1}] \quad (8)$$

where ω_{t-1} refers to the history of the process, i.e., the values up to time t .^j An important characteristic of IRF is that, for linear processes, the IRF is (1) linear in the shock, i.e., $IRF(h, \lambda\delta, \omega_{t-1}) = \lambda IRF(h, \delta, \omega_{t-1})$, (2) symmetric in the shocks, and (3) independent of the history, i.e., $IRF(h, \delta, \omega_{t-1}') = IRF(h, \delta, \omega_{t-1}')$.

^jIn practice, only the last l values, corresponding to the number of lags of the process, are relevant.

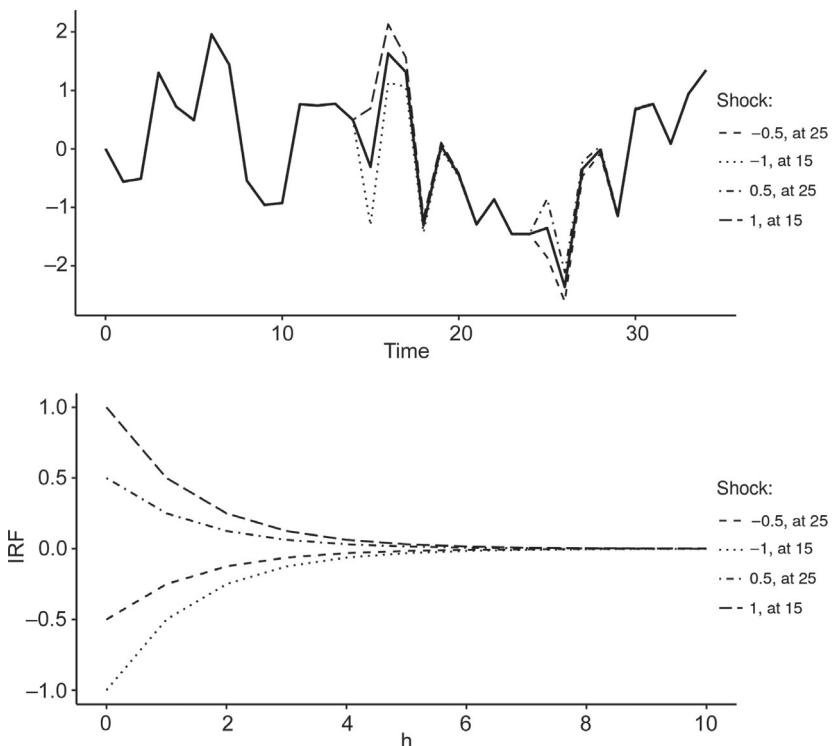


FIG. 1 IRF: illustration for an AR(1).

Fig. 1 illustrates these three properties. In the first panel, **an AR(1) with coefficient 0.5 is generated**. The same series is generated four more times, with shocks 1 and -1 at time 15, and shock 0.5 and -0.5 at time 25. The second panel shows the IRF, computed as the difference between each newly generated series and the original one. One can see that the series are symmetric (a shock of -1 having the same effect than a shock of 1, in absolute value), proportional (a shock of 1 having twice the effect of a shock of 0.5), and finally, that initial conditions (i.e., whether the shock occurs at time 15 or 25) do not matter.

These three nice properties of the IRF function do not hold with nonlinear models, as pointed out by Koop et al. (1996). In the case of a threshold model, **it is clear that initial conditions ω_{t-1} matter**: if the IRF is started in a low or high regime, impacts will be different. Furthermore, the size of the shock δ also matters: is the shock small enough so that the process stays in the same regime, or does it trigger a regime change, and hence has a different impact? Finally, the value of subsequent shocks ϵ_{t+h} matters too: even if the initial shock keeps the process in the same regime, do further shocks trigger a regime change? To see this, Fig. (1) simulates now a SETAR(1, $\rho_L=0.7$, $\rho_H=0.5$, $\theta=-0.5$). In the high regime, the autoregressive coefficient is

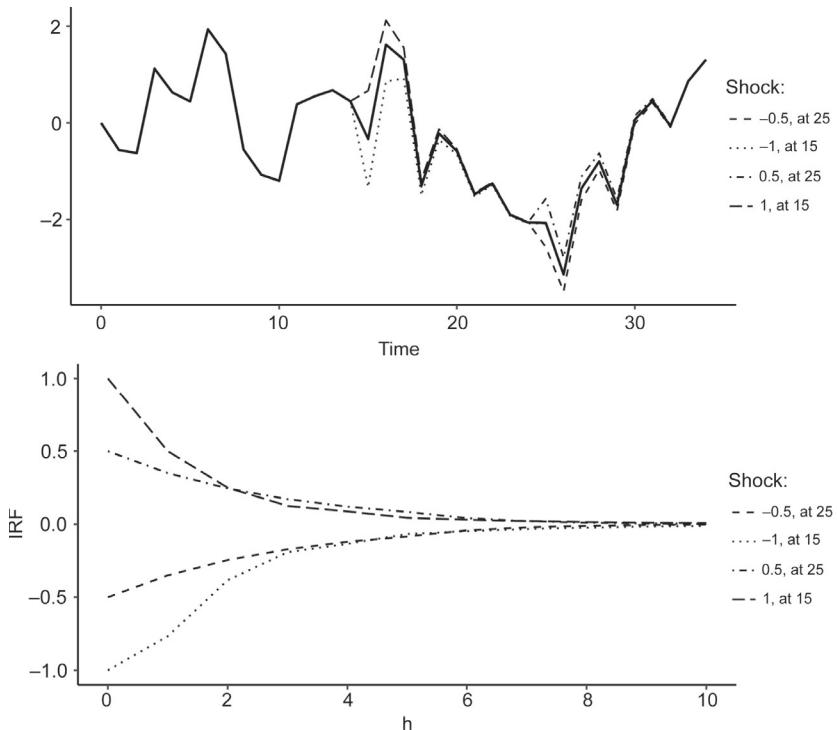


FIG. 2 IRF: illustration for an SETAR(1).

the same as for the AR(1) process above. However, for low values, there is more persistence, $\rho_L = 0.7$. The absence of symmetry of the shocks, as well as the importance of the initial condition, is evident in Fig. (2B).

Based on this, Koop et al. (1996) introduced the **generalized impulse response function (GIRF)**, conditional on shock δ and history ω_{t-1} :

$$GIRF(h, \delta, \omega_{t-1}) = E[y_{t+h} | \epsilon_t = \delta, \omega_{t-1}] - E[y_{t+h} | \omega_{t-1}]. \quad (9)$$

In a GIRF, one is not comparing anymore the shock $\epsilon_t = \delta$ to the (absence of) shock $\epsilon_t = 0$. Instead, the *benchmark* shocks are averaged out. Likewise, subsequent shocks ϵ_{t+h} are not set to 0 as in the IRF, but also averaged out. This is done by Monte-Carlo simulation, using bootstrap draws from the estimated residuals for the innovations.

The $GIRF(h, \delta, \omega_{t-1})$ as defined in Eq. (9) is dependent on the arbitrary shock δ and history ω_{t-1} , as such, it is the realization of the random variable:

$$GIRF(h, \Delta, \Omega_{t-1}) = E[y_{t+h} | \epsilon_t \sim \Delta, \Omega_{t-1}] - E[y_{t+h} | \Omega_{t-1}]. \quad (10)$$

To obtain realizations of $GIRF(h, \delta, \omega_{t-1})$, Koop et al. generate draws from the shocks Δ and starting values Ω_{t-1} . For Δ , they suggest drawing bootstrap samples from the estimated residuals, while for Ω_{t-1} , they suggest using

observed realizations of the raw series. Once a large number of the GIRF realizations is obtained, their density can be plotted for each h . The exercise can be refined in a second step, contrasting the GIRF densities for distinct subset of Ω_{t-1} and Δ , say comparing IRF for initial values in the upper/lower regime, or for positive/negative shocks. This procedure requires unfortunately a large number of simulations: in their empirical application, Koop et al. apparently used the whole history of the series (of size 164), 100 shocks, and 1000 replications for each, amounting to 16 million simulations.

3 The tsDyn package

The **tsDyn** package was initially developed in [Fabio Di Narzo \(2008\)](#), seconded by Jose Luis Aznarte, and focused on nonlinear univariate time series models. It contained an implementation of the SETAR model, as well as alternatives such as the smooth transition autoregressive STAR model, neural networks AR, and additive AR models. [SETAR tests, as well as multivariate linear \(VAR and VECM\) and nonlinear \(TVAR and TVECM\) models were added later on by Matthieu Stigler.](#)

Before describing in detail the functions related to threshold cointegration, some other features are worth describing. In particular, the smooth transition autoregressive (STAR) model is available in functions `lstar()` and `star()`. The STAR model, discussed in the monographs of [Granger and Teräsvirta \(1993\)](#) and [Franses and van Dijk \(2000\)](#), is very similar to the SETAR model, replacing the 1/0 regime indicator function with a smooth function (typically the logistic one). This is meant to describe situations where transitioning from one regime to another is not immediate, but becomes stronger further away from the threshold. Models of cointegration with smooth transitions have also been proposed; see, for example, [Kapetanios et al. \(2006\)](#).

3.1 Linear models: AR, VAR, and VECM in tsDyn

Although package **tsDyn** focuses on nonlinear time series, standard linear models such as AR, VAR, and VECM are also implemented. While for each of these models, more advanced functions can be found in other packages (in particular, `vars` and `urca` for VAR/VECM, see [Pfaff, 2008a,b](#)), the linear functions in **tsDyn** are easy to use together with their nonlinear counterpart.

For univariate models, the standard AR model is available with `linear()`. This function implements the OLS estimator^k and numerous methods are available: `coef()`, `AIC()`, `BIC()`, `predict()`, etc. IRF are available with `irf()`, extending `vars` generic method. Functions `linear`.

^kThe standard R base function `stats::ar()` has five different estimators; on the other side, it lacks basic methods such as `coef()` or `AIC()`.

`sim()` and `linear.boot()` allow, respectively, to simulate an AR from scratch, or resample an estimated AR. To help interpret the lagged coefficients, function `ar_mean()` indicates the long-term mean of a process, and function `charac_root()` returns the roots of the characteristic AR polynomial.¹

Turning to the linear multivariate models, VAR and VECM are available with functions `lineVar()` and `VECM()`. Like for the univariate models, methods such as `AIC()`, `BIC()`, `irf()`, `VAR.sim()/VAR.boot()`, and `VECM.sim()/VECM.boot()` are available. Beside these, a new function `rank.select()` implements a model-selection based approach to simultaneously estimate the cointegrating rank and lag parameters, as suggested by [Gonzalo and Pitarakis \(1998\)](#) and [Aznar and Salvador \(2002\)](#). `rank.test()` implements the standard [Johansen \(1996\)](#) cointegration test, with a somehow more intuitive output than the corresponding `ca.jo()` from `urca`. For more sophisticated analysis, a VECM model from `tsDyn` can be converted into a constrained VAR object `vec2var` from package `vars` using the internal `tsDyn:::vec2var.tsDyn()`. In that way, `vars` functions such as forecast error variance decomposition (`fevd()`), test for serially correlated errors (`serial.test()`) or the ARCH-LM test (`arch.test()`) can be used.

3.2 Univariate models: SETAR

Turning now to univariate threshold models, the main function are `setar()` for a simple interface, and the underlying `selectSETAR()` for more control on the grid search. These allow for pretty general specifications, with one or two thresholds (parameters `nthresh`), multiple lags (`m`), different lags in each regime `mL`, `mM`, `mH`. More general SETAR models can also be estimated, specifying the transition to occur for lag y_{t-j} instead of y_{t-1} by default with argument `thDelay`,^m or specifying an external transition variable `thVar`. As an example, a simple structural break model could be obtained by using `thVar=1:t`. The function `selectSETAR()` will estimate the thresholds parameters using the grid-search described above. The same grid-search can be done for multiple combinations of the lag parameter, allowing to determine simultaneously lags and threshold values based on a AIC or BIC criterion.

¹A stationary AR requires that all values are bigger than 1.

^mIt is important to note that in `tsDyn`, the notation of univariate models differs from the notation of multivariate ones. For univariate model, notation reads as: $y_{t+s} = \rho y_t + \epsilon_t$ for an AR(1), and as $y_{t+s} = \rho y_t^L \mathbb{1}_{y_{t-h} < \gamma} + \rho y_t^H \mathbb{1}_{y_{t-h} \geq \gamma} + \epsilon_t$ for a SETAR(1), that is, the RHS variables are observed at time t . This is different than the notation used in econometrics, where the RHS is at $t-1$. Univariate models in `tsDyn` follow the first notation (RHS is t), while multivariate follow the $t-1$. Importantly, this suggest that a traditional SETAR model with the transition variable being one lag of the dependent is written as `thDelay = 0` if the model is univariate, and `thDelay = 1` for multivariate.

The output from the function `setar()` is a regression model of class `setar` and general class `nlar`. A large number of the methods discussed above for AR/VAR/VECM are available, such as `AIC()`, `BIC()`, `irf()`, `setar.sim()`, etc. New methods particularly relevant in the nonlinear contextⁿ are (1) `regime()`, that returns a vector indicating in which regime the process is at each period, and (2) `GIRF()` for the GIRF, and (3) `irf()` has now the additional argument `regime`, to produce regime-specific IRFs. The method `predict()` takes a new argument `type` for the SETAR, offering different methods to compute confidence intervals, following the procedure discussed in [Franses and van Dijk \(2000, p. 122\)](#).

Testing for threshold effects is available through the function `setarTest()`, which implements [Hansen \(1999\)](#)'s bootstrap procedure. It allows to test three different hypotheses: **1vs2**: Linear AR vs 1 threshold TAR, **1vs3**: Linear AR vs 2 threshold2 TAR, and **2vs3**: 1 threshold TAR vs 2 thresholds TAR. As the two first tests share the same null hypothesis, they are computed together, and obtained with `test="1vs."` The third test is obtained with `test="2vs3."` The downside of the bootstrap procedure is that it requires a very large number of computations: as a grid search is carried for each replication, the procedure is pretty slow. An alternative to this would be to use a model-selection based approach using `AIC()` or `BIC()` to compare a linear and threshold model.

[The Hansen \(1999\) test is meant to be used for stationary process](#). To test for a unit root against a stationary SETAR model function `KapShinTest()` implements the test by [Kapetanios and Shin \(2006\)](#).

3.3 Multivariate models: TVAR and TVECM

Multivariate threshold models are available through the main function `TVAR()` and `TVECM()`. The TVAR model (for threshold VAR) has not been discussed above; in a nutshell, it is a simple extension of the SETAR model to the multivariate case, and can be used for either stationary variables, or nonstationary but noncointegrated ones. The TVECM() model is the one discussed above, which adds an ECT term to the VAR in difference, and switching is based on the lagged value of the ECT term. For estimation of the TVECM, a grid search is conducted over the cointegrating and threshold parameters. In case of a model with three regimes, the grid search over the thresholds parameter is made in the iterative way described above. For the cointegrating parameter, the grid search is made over the confidence intervals derived from the linear cointegrating model, with number of points

ⁿNote that they are also available for linear models, but less relevant in that case.

specified by `ngridBeta=50`. This proved to be too restrictive in practice, so that argument `beta` is provided, with a few options:

exact: Prespecified value.

int: An interval (of length `ngridBeta`) in which to search.

around: A midpoint around which an interval of length `ngridBeta` will be searched for.

When the function `TVECM()` is called, a plot illustrating the grid search is shown, which should help the user refine the search parameters.

Both `TVAR()` and `TVECM()` models return an object of general class `n1Var`, for which the same methods described above are available again (`AIC()`, `BIC()`, `VAR.sim()`, `VAR.boot()`, etc. Likewise, the nonlinear functions `regime()`, `irf()` and `GIRF()` were implemented for `TVAR/TVECM`.

Testing in the `TVECM` can be done using two tests. The first test assumes (linear) cointegration, and tests for threshold cointegration: this is the test of Hansen and Seo (2002), implemented in `TVECM.HTest()`. The second test assumes no cointegration, and has as alternative hypothesis threshold cointegration, this is the Seo (2006) test, implemented in `TVECM.SeoTest()`. These two tests require a bootstrap procedure, and for each step a grid-computation, and hence are rather slow. To alleviate this, parallel processing can be used, with the argument `hpc="foreach"`, see help from package `foreach`.

4 Empirical application

In this section, I illustrate how to run a threshold cointegration analysis, investigating whether the term structure of interest rates theory holds. This dataset has been used in many studies (see, for example, Enders and Siklos, 2001, Hansen and Seo, 2002, Wang et al., 2016), and I show here how to deepen the analysis with the tools available in `tsDyn`.

According to present value models, interest rates of a security with different maturities should be closely related over the long term. Campbell and Shiller (1987) argue that this implies that interest rates with one-period and multiperiod should be cointegrated, with unit vector $(1, -1)$. The data analyzed are interest rates with different maturities, taken from McCulloch and Kwon (1993), who estimate them from the prices of U.S. Treasury securities.^o This data is included as a built-in dataset in the package `tsDyn`, under the name `zeroYld`. The analysis will proceed in three main steps:

1. Analysis of unit roots and linear cointegration
2. Testing, estimation, and interpretation in the univariate residual-based approach
3. Testing, estimation, and interpretation in the multivariate system-based approach

^oData is stored on the website: www.econ.ohio-state.edu/jhm/ts/mcckwon/mccull.htm

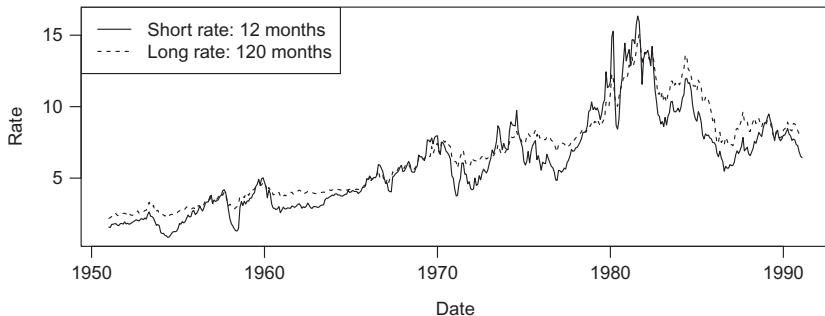
4.1 Unit roots and linear cointegration

The interest rate dataset is contained in the package under the name `zeroYld`, and `zeroYldMeta`, the latter containing in addition the time attributes of the series.

```
> library(tsDyn)
> data(zeroYldMeta)
```

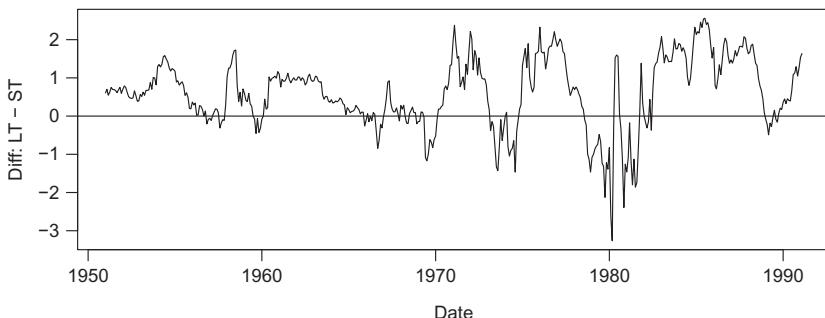
Here, we simply plot the series of the long-term and short-term interest on the government securities:

```
> plot(zeroYldMeta$Date, zeroYldMeta$short.run, type = "l",
+       xlab = "Date", ylab = "Rate")
> lines(zeroYldMeta$Date, zeroYldMeta$long.run, lty = 2)
> legend("topleft", lty = c(1, 2),
+        legend=c("Short rate: 12 months", "Long rate: 120 months"))
```



We see that, in general, the long-term interest rate is higher than the short-term one. There are, however, still 20% of the observations where this relationship is inverted:

```
> diff <- with(zeroYldMeta, long.run - short.run)
> plot(zeroYldMeta$Date, diff, type = "l", xlab = "Date", ylab
+       ="Diff: LT - ST")
> abline(h=0)
```



```
> 100 * mean(diff<0)
[1] 22.61411
```

For the unit root tests, we run the [Elliott et al. \(1996\)](#) test, using the function `ur.ers()`, from package `urca`. Results are shown in the table below. The null hypothesis of a unit root cannot be rejected for the levels of each series, while it is rejected for their difference, suggesting the series are I(1).

| Series | Model | Type | Teststat | 1pct | 5pct | 10pct |
|-----------|-------|--------|----------|-------|-------|-------|
| Short-run | Level | DF-GLS | -0.843 | -2.57 | -1.94 | -1.62 |
| Long-run | Level | DF-GLS | -0.248 | -2.57 | -1.94 | -1.62 |
| Short-run | Diff | DF-GLS | -10.095 | -2.57 | -1.94 | -1.62 |
| Long-run | Diff | DF-GLS | -9.372 | -2.57 | -1.94 | -1.62 |

Turning to the cointegration tests, we use here first the `rank.select()` function form package `tsDyn`. The function does simultaneous selection of cointegrating rank and lags using information criteria:

```
> rank.select(zeroYldMeta[, c("long.run", "short.run")])

Best AIC: rank= 2 lag= 8
Best BIC: rank= 1 lag= 1
Best HQ : rank= 1 lag= 1
```

Results are somewhat mixed: the BIC and the Hannan-Quinn criteria suggest a cointegrated system (rank 1) and a single lag, while the AIC criterion suggests actually a stationary VAR with eight lags. The former result is at odds with the initial finding of unit roots for each series, so we will follow the results from the BIC. More formal testing can be conducted using the standard Johansen test, function `rank.test()`, which can be called after estimating a VECM:

```
> vecm_simple<- VECM(zeroYldMeta[, c("long.run", "short.run")], r=1,
+                         lag =1, estim = "ML")
> rank.test(vecm_simple)
```

Rank selected: 1 (first eigen test with pval above 5 %: 10 %)

The Johansen trace test suggests that the data are cointegrated. Alternatives to using this function could be the similar [Johansen \(1996\)](#) `ca.jo()` or the [Phillips and Ouliaris \(1990\)](#) test `ca.po()`, both in package `urca`.

Once linear cointegration is established, we turn to examining the results, printing the object directly, or using the `summary` method for a few more details:

```
> summary(vecm_simple)
#####
###Model VECM
#####
Full sample size: 482      End sample size: 480
Number of variables: 2      Number of estimated slope parameters 8
AIC -2155.763      BIC -2118.199      SSR 176.3897
Cointegrating vector (estimated by ML):
  long.run short.run
r1      1 -1.022065

          ECT           Intercept
Equation long.run -0.0116(0.0153) 0.0169(0.0158)
Equation short.run 0.0888(0.0262)*** -0.0362(0.0271)
          long.run -1      short.run -1
Equation long.run 0.0480(0.0682) 0.0116(0.0392)
Equation short.run 0.3254(0.1171)** 0.0506(0.0673)
```

From this, we see that it is mainly the short-term interest rate that reacts to the deviations (the ECT term). Looking at the cointegrating vector, it is very close to 1, as predicted by theory. We can use this value instead. For simple restrictions on the cointegrating parameters, `VECM()` has the argument `beta`, which allows to include a user-specified value. For an informal test, we will have a look at how the restriction affects the log-likelihood:

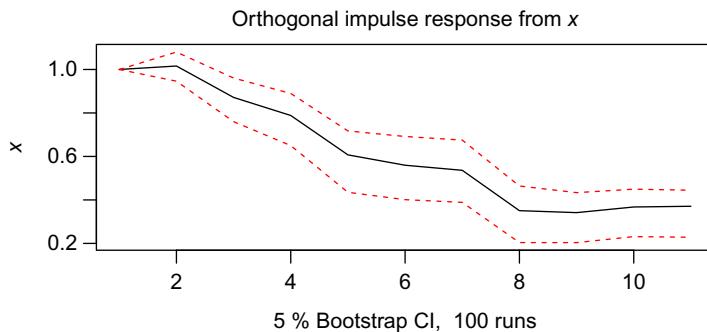
```
> vecm_rest <- VECM(zeroYldMeta[, c("long.run", "short.run")], r=1,
+                      lag = 1, estim = "ML", beta = c(1, -1))
> ll_restricted = logLik(vecm_rest), ll_unrestricted =
logLik(vecm_simple))

ll_restricted ll_unrestricted
-275.3940      -275.2995
```

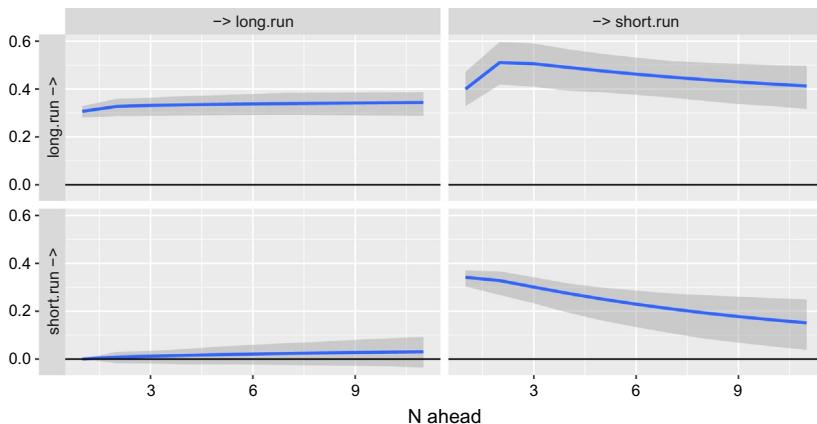
The change in the log-likelihood seems to be very small, suggesting that there is no harm in imposing the value of 1. For more formal test on the cointegrating vector, the reader is referred to the function `blrtest()` in package `urca`.

We turn now to investigating the dynamics of the model, looking at the IRF, called with `irf()`. In a somehow unusual way, we look first at the IRF for the ECT term itself.

```
> ect <- with(zeroYldMeta, long.run - short.run)
> ect_ar <- linear(ect, m = 8)
> ect_ar_irf <- irf(ect_ar, n.ahead = 10)
> plot(ect_ar_irf)
```



We then use the same `irf()` function, this time on the VECM output:



From the IRF analysis of the error correction term, we see that the ECT shows a rather high persistence, as shocks do not get close to 0 after 10 time periods. Indeed, one would need to set `irf(ect_ar, n.ahead=50)` to see the IRF get close to 0. The IRF analysis based on the VECM tells us that the long-run variable has some pretty high persistence, while the short-run one decays more rapidly. Looking at their cross-relationships, we see that the short-run variable responds to shocks from the long-run one, while the opposite does not hold. This confirms the results we saw while looking at the speed adjustment coefficients in the VECM.

4.2 Threshold cointegration in the univariate residual-based approach

In the previous section, we saw that the short and long-run interest rates were cointegrated, with vector $(1, -1)$. This finding will make the analysis much

easier: to test for threshold cointegration, we can simply test for linear cointegration vs threshold cointegration. Furthermore, as the vector can be assumed to be known, we can use a simple linear AR vs threshold AR (SETAR) test.

We start here with the information-based approach, using AIC and BIC measures, as advocated by [Gonzalo and Pitarakis \(2002\)](#). This is simply done by computing the AR model (using `linear(ect, m=1)`) and the SETAR (`setar(ect, m=1, nthresh=1)`) for various values of lag m and number of thresholds nthresh , then using functions `AIC()` and `BIC()`. Results are shown in the table below:

| | AIC (l1) | AIC (l2) | AIC (l8) | BIC (l1) | BIC (l2) | BIC (l8) |
|----------|----------|----------|----------|----------|----------|----------|
| AR | -964.39 | -967.17 | -978.47 | -956.03 | -954.64 | -940.86 |
| SETAR(1) | -978.01 | -989.06 | -1001.87 | -957.12 | -959.81 | -922.49 |
| SETAR(2) | -977.62 | -989.67 | -999.03 | -944.20 | -943.71 | -877.87 |

For almost every lag, the SETAR with one threshold is selected by the AIC and BIC over the AR, and over the SETAR(2). There is a slight conflict between the results from the AIC and BIC, the AIC selecting a SETAR(1) with eight lags, while the BIC selects the same model, but with two lags only. The fact that the BIC selects a more parsimonious model than the AIC is an expected outcome, as the BIC puts a stronger penalty on the number of parameters. To avoid overfitting, we will proceed with a SETAR(1) with two lags, following the BIC criterion.

To confirm these results, we use the `setarTest()` function, which implements the bootstrap version of [Hansen \(1999\)](#) test. The first version tests an AR vs a SETAR(1) or SETAR(2):

```
> setarTest(ect, m = 2, nboot = 1000)

Test of linearity against setar(2) and setar(3)

      Test  Pval
1vs2  53.96306   0
1vs3  70.83488   0
```

The null of the linear model is rejected both in favor of the SETAR(1) and SETAR(2). To decide between these two models, we run the same test, this time with the SETAR(1) as the null hypothesis, and the SETAR(2) as the alternative, using `test="2vs3"` instead of the default value `test="1vs"`:

```
> setarTest(ect, m = 2, nboot = 1000, test = "2vs3")

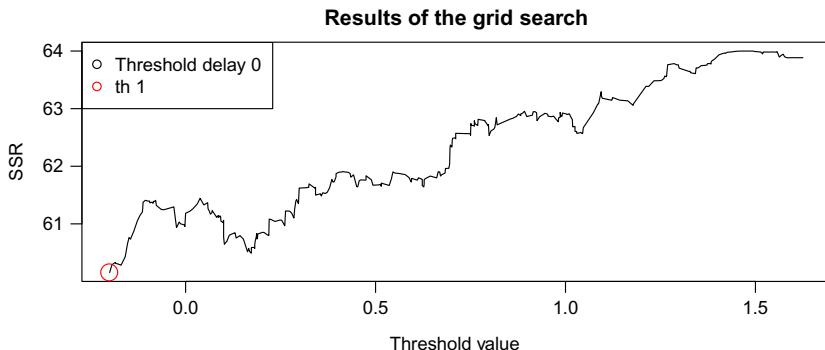
Test of setar(2) against setar(3)

      Test  Pval
2vs3 15.16673 0.065
```

The test barely rejects the null hypothesis, at 6.5%. Following the results obtained with the AIC/BIC criteria, we opt here for the SETAR(1) model.

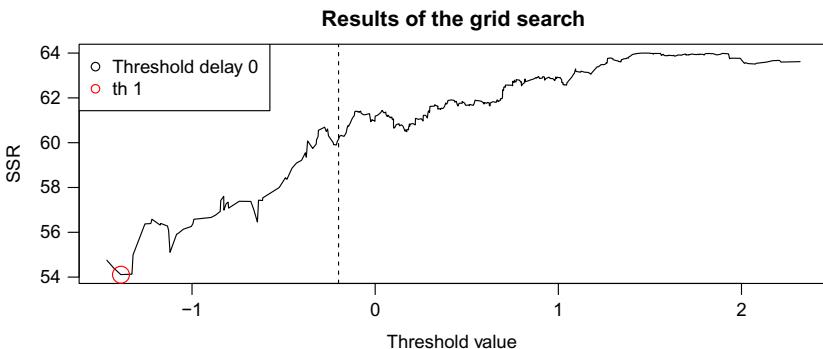
We turn now to estimating the SETAR(1) model itself. This can be done directly using the functions `setar()` or `selectSETAR()`. The latter function allows to run the grid-search for multiple values of the lags, and returns also the full results from the grid search. This is useful for illustrative purpose, so we use it here:

```
> grid_full <- selectSETAR(ect, m=2, trim = 0.15,
+                               trace = FALSE, plot = FALSE,
+                               criterion = "SSR")
> plot(grid_full, type = "l")
```



The figure shows the values of the concentrated sum of squares $SSR(\gamma) = SSR(\hat{\theta}(\gamma), \gamma)$ (see Eq. 7), for each γ contained in $[\gamma_{15\%}, \gamma_{85\%}]$, where $\gamma_{15\%}$ corresponds to the 15% quantile of the threshold variable, i.e., the ECT. The rationale for restricting the search to values in the interval $[\gamma_{15\%}, \gamma_{85\%}]$ is to ensure a minimum of observations in each regime, for statistical accuracy. This minimum percentage of observations can be adjusted using the `trim` argument, which by default is 15%. The threshold parameter is the minimizer $\hat{\gamma} = \arg \min_{\gamma} SSR(\gamma)$, and is shown in red in the figure. It turns out that the minimizer is here at the boundary of the interval, suggesting that the arbitrary value of 15% is binding, and that the minimum might be actually below the 15% quantile. We rerun then the same function with a much lower `trim` value:

```
> grid_full2 <- selectSETAR(ect, m=2, trim = 0.015,
+                               trace = FALSE, plot = FALSE, criterion =
+                               "SSR")
> plot(grid_full2, type = "l")
> abline(v = getTh(grid_full)$th, lty = 2)
```



We see that extending the search over a larger interval leads to a minimiser that is much lower, and that the value found in the first step (shown with the dotted line) does not seem to be a minimizer. We are, however, facing now a dilemma: if we were to use the threshold parameter as found in the more exhaustive search, we would end up with a regime containing 2.5% of the observations (i.e., 11 points). This is certainly not enough to obtain reliable estimates of the AR coefficients. We address this dilemma somewhat arbitrarily, by setting the `trim` parameter at 8%, which corresponds to the clear spike seen at -0.6 , and have now 43 observations in the low regime. Estimation of the SETAR model itself is obtained with the function `setar()`:

```
> set_12 <- setar(ect, m = 2, trim = 0.08)
> set_12
```

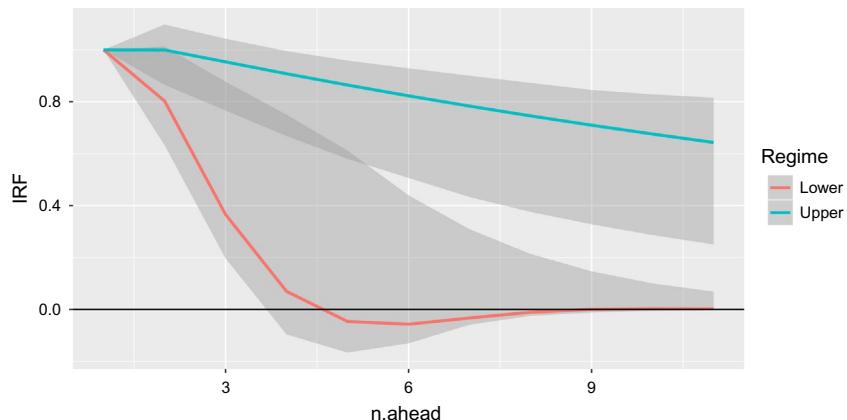
```
Non linear autoregressive model
SETAR model ( 2 regimes)
Coefficients:
Low regime:
  const.L  phiL.1  phiL.2
-0.9789932 0.3844136 -0.3204736
High regime:
  const.H  phiH.1  phiH.2
0.03933391 1.06928188 -0.13467791
Threshold:
-Variable: Z(t) = + (1) X(t)+ (0)X(t-1)
-Value: -0.643
Proportion of points in low regime: 9.17% High regime: 90.83%
```

We hence obtained two regimes, one containing close to 90% of the observations, and one with 10% of the observations, describing the situation when the

ECT is negative. Remembering that the ECT represents here the difference between the long-term and the short-term interest rates, a negative ECT corresponds to the so-called *inverted yield curve*, where less risky assets (short-term ones) give actually higher yields than more risky ones. From now on, we will refer to the two regimes as the *normal* one (positive yield curve, or slight inversion) and the *inverted-yield* one. Direct comparison of the coefficients is difficult as soon as there is more than one lag. We resort first to the regime-specific IRFs:

```
> irf_th1_12_L <- irf(ect_set_th1_12, regime = "L")
> irf_th1_12_H <- irf(ect_set_th1_12, regime = "H")
```

Combining both IRF in a single graphs shows an important difference in the persistence profile: while shocks in the *normal* regime have a rather high persistence (as we saw in the previous section looking at the linear model),



shocks in the inverted-yield regime dissipate much faster.

While they deliver useful information, the linear regime-specific IRF functions have the drawback that they do not take into account possible regime switching. We can use the function `regime()` to extract an indicator for the regime, and compute what is the average, minimum or maximum duration in each regime (using R's `rle()` function):

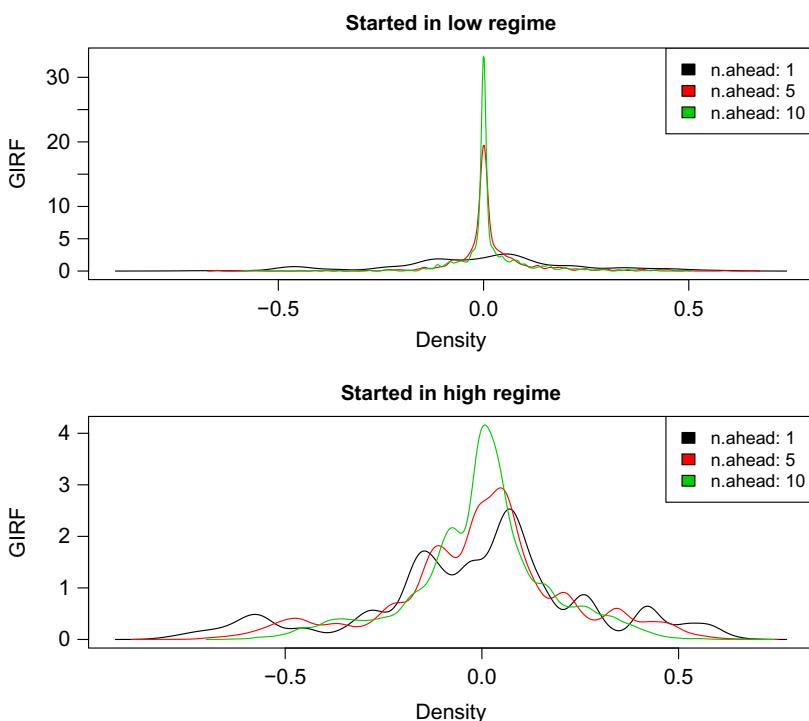
```
> reg_set_12 <- regime(set_12, initVal = FALSE, timeAttr = FALSE)
> rl <- rle(reg_set_12)
> aggregate(rl$lengths, by = list(regime = rl$values),
+             function(x) c(mean=mean(x), max=max(x), min=min(x)))
   regime      x.mean      x.max      x.min
1       1  4.888889  9.000000  1.000000
2       2 43.600000 187.000000  1.000000
```

We see that the process does not “stay” for long in the *inverted-yield* regime: the average duration is just five periods, and the longest duration is nine periods. This suggests that the IRF shown above is not very representative of the process: it actually never happens that the process stays in the lower regime for 10 periods. To investigate the impact of shocks taking into account possible regime-switching, we use now the GIRF method, with the `GIRF()` function:

```
> set_GIRF <- GIRF(object=ect_set_th1_12, n.hist = 200, n.shock =50)
```

The output of `GIRF()` is a simple `data.frame`, containing the GIRF replications, together with the shock and history considered for each replication. We will here look separately at simulations that started in the *normal* vs *inverted-yield* regimes. As suggested by [Koop et al. \(1996\)](#), we look at the density for distinct `n.ahead` values: the persistence of a process can be judged by how slowly the density of the GIRF concentrate around zero for higher `n.ahead` values.

```
> girf_high <- subset(set_GIRF, hist_x1_11>getTh(set_12))
> girf_low <- subset(set_GIRF, hist_x1_11 <=getTh(set_12))
> par(mfrow = c(2, 1))
> plot(girf_low, main = "Started in low regime")
> plot(girf_high, main = "Started in high regime")
```



```
> par(mfrow = c(1, 1))
```

We see that persistence is indeed much faster for shocks starting in the *inverted-yield* regime: despite having much larger initial response (see black curve for `n.ahead=1`), the response after five time periods is already close to zero for most of the shocks. On the other side, for shocks originating in the *normal* regime, the initial impact is narrower, but even after 10 periods, is still not fully absorbed.

4.3 Threshold cointegration in the multivariate system-based approach

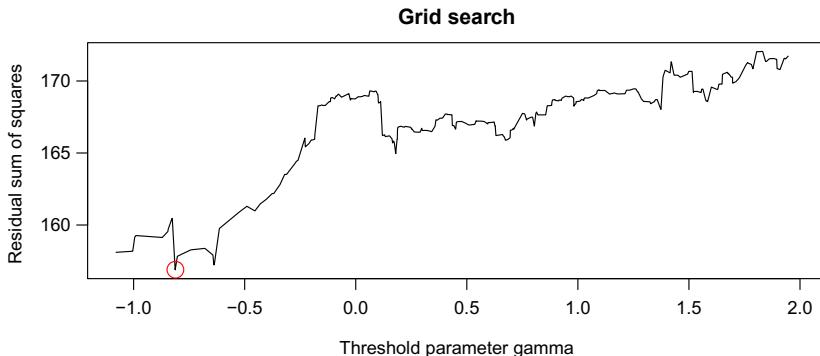
In the previous Section, we investigated the dynamics of the interest rates using the residual-based univariate approach. We found that the error correction term was following a SETAR(1) model, distinguishing between a normal regime (containing 90% of the observations), and a special regime, that corresponds to a situation of strong *yield-inversion*. In that regime, persistence is much lower and shocks dissipate much faster. Economically, this makes sense, as the yield-inverted regime is an unusual situation, that is not expected to last very long.

The next question to ask is which variable is responsible for this change in persistence. This can be answered within the system-based approach, with a threshold VECM (TVECM). To start with, we test once more for threshold cointegration. As we found linear cointegration to hold, we run the [Hansen and Seo \(2002\)](#) test, which tests linear cointegration vs threshold cointegration. The function `TVECM.HStest()` implements the test:

```
> zero_hstest <- TVECM.HStest(data = zeroYldMeta[, c("long.run",
  "short.run")],
  nboot = 500, lag = 2)
## Test of linear versus threshold cointegration of Hansen and
## Seo (2002) ##
Test Statistic: 20.952 (Maximized for threshold value: 0.173 )
P-Value: 0.046 ( Fixed regressor bootstrap )
```

The null hypothesis of linear cointegration is rejected at 4% against the threshold cointegration alternative, confirming the results we obtained in the residual-based approach. We proceed then to estimating the TVECM itself, using the `TVECM()` function. The `TVECM()` function runs a two-dimensional grid search, over the cointegrating and threshold parameter. Arguments `ngridBeta` and `ngridTh` indicate how many points should be considered for each parameter. Arguments `th1` and `beta` allow to refine, or extend the search, by using an prespecified value (`exact`), an interval to search over (`int`), or a point to search around (`around`). Here, we will simply use the cointegrating value of 1 that we found in the linear step.

```
> tvec_b1 <- TVECM(zeroYldMeta[, c("long.run", "short.run")],
+ lag = 2,
+ ngridTh = 300, beta = list(exact = 1))
```



```
> tvec_b1
Model TVECM with 1 thresholds

$Bdown
          ECT      Const long.run t -1 short.run t -1
Equation long.run -0.0763864 0.2211566   0.1123809 -0.20666272
Equation short.run 0.9746974 1.2234335   0.5555664  0.06697354
                           long.run t -2 short.run t -2
Equation long.run     -0.4395419    -0.10741548
Equation short.run    -0.7689766    0.08647546

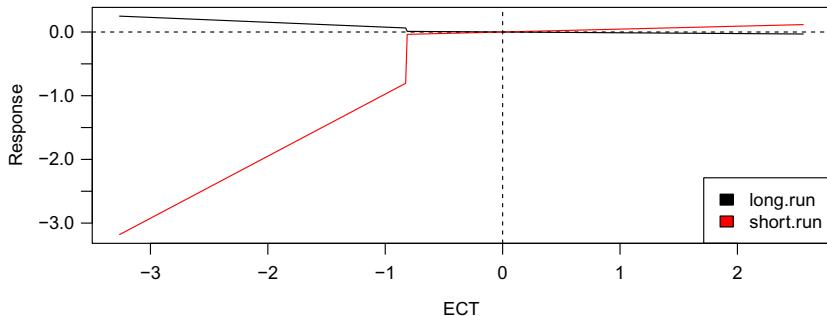
$Bup
          ECT      Const long.runt -1 short.runt -1
Equation long.run -0.01224638 0.01531765 -0.03253936   0.0781046
Equation short.run 0.04539740 -0.01402965  0.14929464   0.1672832
                           long.run t -2 short.run t -2
Equation long.run    0.007451146   -0.01787635
Equation short.run   0.093122884   -0.04649894

Threshold value[1] -0.813
```

As we provided a fixed value for the cointegrating parameter, only a one-dimensional search over the threshold parameter is run, and the default plot method shows the resulting grid. The SSR profile looks very similar to the one obtained in the univariate approach, and the estimated threshold found with the TVECM is -0.813 , close to the value of -0.643 found in the SETAR. Looking at the speed adjustment coefficient for the ECT, we see that

the short-term variable responds much faster in the *inverted-yield* regime. Function `plot_ECT()` shows the regime-specific adjustment to the ECT:

```
> plot_ECT(tvec_b1, legend.location = "bottomright")
```

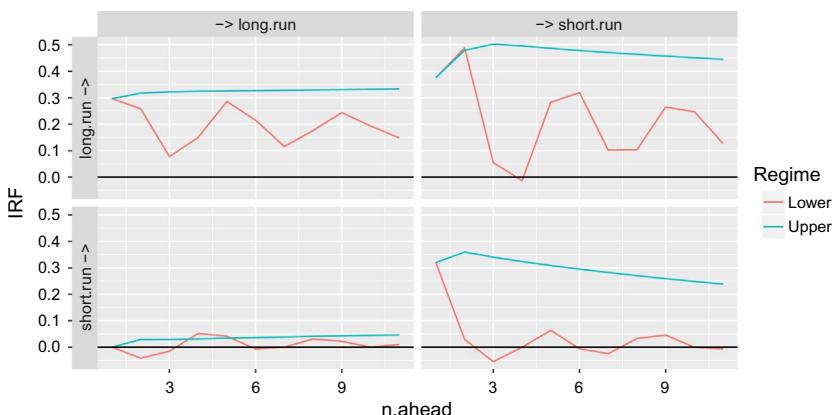


The difference in the speed adjustment between regimes is striking. In fact, using `summary(tvec_b1)` reveals that the speed adjustment parameters in the *normal* regime are now not significantly different from zero.

We use now the regime-specific linear IRF functions. The bootstrap method to obtain confidence intervals is rather slow, as it involves reestimating the grid at each step, so we set `boot=FALSE` here:

```
> tvecm_irf_L <- irf(tvec_b1, regime = "L", boot = FALSE, n.ahead = 10)
> tvecm_irf_H <- irf(tvec_b1, regime = "H", boot = FALSE, n.ahead = 10)
```

Combining the two outputs, we obtain the following figure:

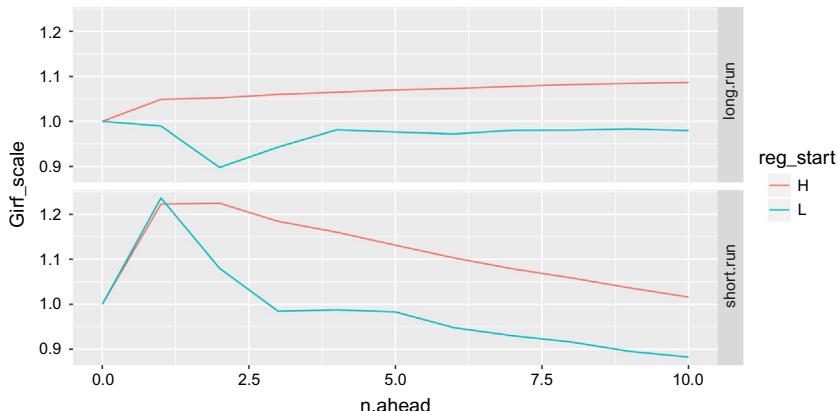


The IRF in the lower regime exhibit a cyclical pattern, which is probably due to the negative autocorrelation found in the coefficients of the lagged variables. We see that most of the changing dynamics are due to the short-run interest rate: it responds much faster to shocks in the long-run variable, and especially faster in the *inverted-yield* regime.

Finally, we run the GIRF on the TVECM output, using the full history, together with 200 innovations. This is a heavy operation: computation takes a while (even if we use only $R=20$ inner simulations for each of the $480 * 200$ simulation), and we end-up with a large dataset.

```
> GIRF_multi <- GIRF(tvec_b1, n.hist = 480, n.shock = 100)
```

To show the output of this multivariate nonlinear GIRF, I suggest here to look at the average standardized shock, depending on whether the system was previously in the low or high regime. To do so, I standardize every GIRF path, setting the initial shock to be 1.^p I then average the standardized paths for each variable, and for each regime.^q



This indicates that an average shock in the long-run variable has a stronger and more persistent impact if it originates in the high-regime than the *inverted-yield* one. The same effect is to be seen in the short-run variable. The latter shows first a strong initial reaction (as can be seen in the ECT coefficient close to 1), but decays faster later on.

^pMore precisely, for a given simulation i , I use: $\text{GIRF}(i, \text{n.ahead})/\text{GIRF}(i, \text{n.ahead} = 1)$, setting the initial shock to be 1.

^qAveraging was done using the median rather than the mean, as some GIRF path showed some extreme behavior under the low regime. Setting a higher value of R would have probably helped.

Summarizing the contribution of the system-based TVECM compared to the SETAR, we saw here that most of the asymmetry in the regime came from the short-run variable. Not only does the short-run variable respond usually faster compared to the long-run variable to changes in the ECT, but it does so even faster in the *inverted-yield* regimes. This explains the phenomenon we observed in the univariate analysis that shocks in the inverted-yield regime were resorbed quickly.

5 Conclusion

The threshold cointegration model is a very powerful extension of the concept of cointegration, allowing to take into account economic phenomena such as asymmetry or noninstantaneous adjustment due to transaction costs. The model has been indeed used and applied in a wide variety of settings such as agricultural markets, oil prices, interest rates etc. In this chapter, I surveyed the seminal papers and their later developments, describing the various estimators and tests found in the literature. I described also how to interpret the results, using various tools such as an “error correction plot,” regime-specific IRF and their generalized version, the GIRF. In the empirical application, I showed how each of these tools could be used to investigate the dynamics of long-run and short-run interest rates.

The power of the threshold cointegration model comes, however, at a certain cost, as estimation and testing for the model is involved, both at the theoretical and computational levels. It is my hope that package **tsDyn** will help the user alleviate some of these difficulties. Some readers might be obfuscated by the sometimes arbitrary decisions taken while modeling the empirical application. The aim of this exercise is, however, not so much to seek to convince the reader about the correctness of the analysis undertaken here, but rather to highlight some of the challenges and dilemmas a user will typically face.

Acknowledgment

I am grateful to Charlotte Ambrozek and Jessica Rudder for their helpful editing assistance, and to Aaron Smith for his valuable support.

References

- Al-Abri, A.S., Goodwin, B.K., 2009. Re-examining the exchange rate pass-through into import prices using non-linear estimation techniques: threshold cointegration. *Int. Rev. Econ. Financ.* 18 (1), 142–161. URL, <http://ideas.repec.org/a/eee/reveco/v18y2009i1p142-161.html>.
- Andrews, D.W.K., 1993. Tests for parameter instability and structural change with unknown change point. *Econometrica* 61 (4), 821–856. URL, <http://www.jstor.org/stable/2951764>.
- Andrews, D.W.K., Ploberger, W., 1994. Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62 (6), 1383–1414. URL, <http://ideas.repec.org/a/ecm/emetrp/v62y1994i6p1383-1414.html>.

- Aznar, A., Salvador, M., 2002. Selecting the rank of the cointegration space and the form of the intercept using an information criterion. *Econom. Theor.* 18 (04), 926–947. URL, http://ideas.repec.org/a/cup/etheor/v18y2002i04p926-947_18.html.
- Bai, J., Perron, P., 2003. Computation and analysis of multiple structural change models. *J. Appl. Econom.* 18 (1), 1–22. URL, <http://www.jstor.org/stable/30035185>.
- Balke, N.S., Fomby, T.B., 1997. Threshold cointegration. *Int. Econ. Rev.* 38 (3), 627–645. URL, <http://ideas.repec.org/a/ier/iecrev/v38y1997i3p627-45.html>.
- Bec, F., Salem, M.B., Carrasco, M., 2004. Tests for unit-root versus threshold specification with an application to the purchasing power parity relationship. *J. Bus. Econ. Stat.* 22, 382–395. URL, <http://ideas.repec.org/a/bes/jnlbes/v22y2004p382-395.html>.
- Bec, F., Guay, A., Guerre, E., 2008. Adaptive consistent unit-root tests based on autoregressive threshold model. *J. Econ.* 142 (1), 94–133. URL, <http://ideas.repec.org/a/eee/econom/v142y2008i1p94-133.html>.
- Campbell, J.Y., Shiller, R.J., 1987. Cointegration and tests of present value models. *J. Polit. Econ.* 95 (5), 1062–1088. URL, <https://doi.org/10.1086/261502>.
- Chan, K., 1993. Consistency and limiting distribution of the least squares estimation of a threshold autoregressive model. *Ann. Stat.* 21, 520–533.
- Chan, K., Tong, H., 1990. On likelihood ratio tests for threshold autoregression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 52 (3), 469–476.
- Chan, K., Petruccielli, J., Woolford, S., Tong, H., 1985. A multiple threshold AR(1) model. *J. Appl. Probab.* 22 (2), 267–279.
- Damanjani, R., Yang, B.Z., 1998. Price rigidity and asymmetric price adjustment in a repeated oligopoly. *J. Inst. Theor. Econ.* 154 (4), 659. [http://ideas.repec.org/a/mhr/jinst/urnsici0932-4569\(199812\)1544_659praapa_2.0.tx_2-.html](http://ideas.repec.org/a/mhr/jinst/urnsici0932-4569(199812)1544_659praapa_2.0.tx_2-.html).
- Davies, R., 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–254.
- Davies, R., 1987. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74, 33–43.
- Dutta, S., Bergen, M., Levy, D., Venable, R., 1999. Menu costs, posted prices, and multiproduct retailers. *J. Money Credit Bank.* 31 (4), 683–703.
- Elliott, G., Rothenberg, T., Stock, J., 1996. Efficient tests for an autoregressive unit root. *Econometrica* 64 (4), 813–836.
- El-Shagi, M., 2011. An evolutionary algorithm for the estimation of threshold vector error correction models. *Int. Econ. Econ. Policy* 8 (4), 341–362. URL, <https://ideas.repec.org/a/kap/iecepo/v8y2011i4p341-362.html>.
- Enders, W., Siklos, P.L., 2001. Cointegration and threshold adjustment. *J. Bus. Econ. Stat.* 19 (2), 166–176. URL, <http://ideas.repec.org/a/bes/jnlbes/v19y2001i2p166-76.html>.
- Engle, R.F., Granger, C., 1987. Co-integration and error correction: representation, estimation and testing. *Econometrica* 55 (2), 251–276.
- Fabio Di Narzo, A., 2008. Nonlinear Autoregressive Time Series Models in R Using tsDyn Version 0.7. URL, <https://cran.r-project.org/package=tsDyn/vignettes/tsDyn.pdf>.
- Franses, P.H., van Dijk, D., 2000. Nonlinear Time Series Models in Empirical Finance. Cambridge University Press.
- Ghassan, H.B., Banerjee, P.K., 2015. A threshold cointegration analysis of asymmetric adjustment of OPEC and non-OPEC monthly crude oil prices. *Empir. Econ.* 49 (1), 305–323. URL, <https://doi.org/10.1007/s00181-014-0848-0>.
- Gonzalo, J., Pitarakis, J.-Y., 1998. Specification via model selection in vector error correction models. *Econ. Lett.* 60 (3), 321–328. URL, <http://www.sciencedirect.com/science/article/B6V84-3TX5F4P-C/2/04b30495869fc966437dd83b634d9b79>.

- Gonzalo, J., Pitarakis, J.-Y., 2002. Estimation and model selection based inference in single and multiple threshold models. *J. Econ.* 110 (2), 319–352. URL, <http://www.sciencedirect.com/science/article/B6VC0-4684TG7-1/2/624fe6c57473ed0ed50a2a98633499ec>.
- Gonzalo, J., Pitarakis, J., 2006a. Threshold effects in multivariate error correction models. In: Mills, T.C., Patterson, K. (Eds.), *Palgrave Handbook of Econometrics. Econometric Theory*, vol. 1. Palgrave MacMillan, pp. 578–609.
- Gonzalo, J., Pitarakis, J.-Y., 2006b. Threshold effects in cointegrating relationships. *Oxf. Bull. Econ. Stat.* 68 (s1), 813–833. URL, <http://ideas.repec.org/a/bla/obuest/v68y2006is1p813-833.html>.
- Gonzalo, J., Wolf, M., 2005. Subsampling inference in threshold autoregressive models. *J. Econ.* 127 (2), 201–224. URL, <http://ideas.repec.org/a/eee/econom/v127y2005i2p201-224.html>.
- Gouveia, P., Rodrigues, P., 2004. Threshold cointegration and the PPP hypothesis. *J. Appl. Stat.* 31 (1), 115–127. URL, <http://ideas.repec.org/a/taf/japsta/v31y2004i1p115-127.html>.
- Granger, C.W.J., Lee, T.H., 1989. Investigation of production, sales and inventory relationships using multicointegration and non-symmetric error correction models. *J. Appl. Econom.* 4, 145–159.
- Granger, C., Newbold, P., 1974. Spurious regressions in econometrics. *J. Econ.* 2, 111–120.
- Granger, C., Teräsvirta, T., 1993. *Modelling Nonlinear Economic Relationships*. Oxford University Press, New York.
- Hammoudeh, S.M., Ewing, B.T., Thompson, M.A., 2008. Threshold cointegration analysis of crude oil benchmarks. *Energy J.* 29 (4), 79–95. URL, <http://www.jstor.org/stable/41323182>.
- Hansen, B.E., 1996. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* 64 (2), 413–430. URL, <http://ideas.repec.org/a/ecm/emetrp/v64y1996i2p413-30.html>.
- Hansen, B., 1999. Testing for linearity. *J. Econ. Surv.* 13 (5), 551–576.
- Hansen, B.E., 2000. Sample splitting and threshold estimation. *Econometrica* 68 (3), 575–604. URL, <http://ideas.repec.org/a/ecm/emetrp/v68y2000i3p575-604.html>.
- Hansen, B., Seo, B., 2002. Testing for two-regime threshold cointegration in vector error-correction models. *J. Econ.* 110, 293–318.
- Heimonen, K., 2006. Nonlinear adjustment in PPP evidence from threshold cointegration. *Empir. Econ.* 31 (2), 479–495. URL, <http://ideas.repec.org/a/spr/empeco/v31y2006i2p479-495.html>.
- Horvath, M.T., Watson, M.W., 1995. Testing for cointegration when some of the cointegrating vectors are prespecified. *Econom. Theor.* 11 (05), 984–1014. URL, http://ideas.repec.org/a/cup/etheor/v11y1995i05p984-1014_00.html.
- Jawadi, F., Million, N., Arouri, M.E.H., 2009. Stock market integration in the latin American markets: further evidence from nonlinear modeling. *Econ. Bull.* 29 (1), 162–168.
- Johansen, S., 1988. Statistical analysis of cointegration vectors. *J. Econ. Dyn. Control.* 12, 231–254.
- Johansen, S., 1996. *Likelihood-Based Inference in Cointegrated Vector Autoregresive Models*. Oxford University Press.
- Kapetanios, G., Shin, Y., 2006. Unit root tests in three-regime setar models. *Econ. J.* 9 (2), 252–278. URL, <http://ideas.repec.org/a/ect/emjrnl/v9y2006i2p252-278.html>.
- Kapetanios, G., Shin, Y., Snell, A., 2006. Testing for cointegration in nonlinear smooth transition error correction models. *Econom. Theor.* 22 (2), 279–303. URL, <http://www.jstor.org/stable/4093226>.
- Killick, R., Fearnhead, P., Eckley, I.A., 2012. Optimal detection of changepoints with a linear computational cost. *J. Am. Stat. Assoc.* 107 (500), 1590–1598. URL, <https://doi.org/10.1080/01621459.2012.737745>.

- Koop, G., Pesaran, M.H., Potter, S.M., 1996. Impulse response analysis in nonlinear multivariate models. *J. Econ.* 74 (1), 119–147. URL, <http://ideas.repec.org/a/eee/econom/v74y1996i1p119-147.html>.
- Krishnakumar, J., Neto, D., 2008. Testing Uncovered Interest Rate Parity and Term Structure Using Multivariate Threshold Cointegration. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 191–210. URL, https://doi.org/10.1007/978-3-540-77958-2_10.
- Krishnakumar, J., Neto, D., 2015. Testing for the cointegration rank in threshold cointegrated systems with multiple cointegrating relationships. *Stat.Methodol.* 26, 84–102. URL, <http://www.sciencedirect.com/science/article/pii/S1572312715000301>.
- Levy, D., Bergen, M., Dutta, S., Venable, R., 1997. The magnitude of menu costs: direct evidence from large U. S. supermarket chains. *Q. J. Econ.* 112 (3), 791–825.
- Li, J., 2017. System-equation ADL test for threshold cointegration with an application to the term structure of interest rates. *Oxf. Bull. Econ. Stat.* 79 (1), 1–24. URL, <https://onlinelibrary.wiley.com/doi/abs/10.1111/obes.12123>.
- Li, J., Lee, J., 2010. ADL tests for threshold cointegration. *J. Time Ser. Anal.* 31 (4), 241–254. URL, <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9892.2010.00659.x>.
- Li, D., Ling, S., 2012. On the least squares estimation of multiple-regime threshold autoregressive models. *J. Econ.* 167 (1), 240–253. URL, <http://www.sciencedirect.com/science/article/pii/S0304407611002685>.
- Lo, M.C., Zivot, E., 2001. Threshold cointegration and nonlinear adjustment to the law of one price. *Macroecon. Dyn.* 5, 533–576.
- Maki, D., 2009. Tests for a unit root using three-regime tar models: power comparison and some applications. *Econ. Rev.* 28 (4), 335–363. URL, <http://ideas.repec.org/a/taf/emetrv/v28y2009i4p335-363.html>.
- McCulloch, J.H., Kwon, H.-C., 1993. US Term Structure Data, 1947–1991. Tech. Rep. 93–6, Ohio State University Working Paper. March.
- Meyer, J., 2004. Measuring market integration in the presence of transaction costs—a threshold vector error correction approach. *Agric. Econ.* 31 (2–3), 327–334.
- Million, N., 2004. Central bank's interventions and the fisher hypothesis: a threshold cointegration investigation. *Ecol. Model.* 21 (6), 1051–1064. URL, <http://ideas.repec.org/a/eee/ecmode/v21y2004i6p1051-1064.html>.
- Nelson, Plosser, 1982. Trends and random walks in macroeconomic time series: some evidence and implications. *J. Monet. Econ.* 10 (2), 139–162.
- Petrucci, J., Davies, N., 1986. A portmanteau test for self-exciting threshold autoregressive-type nonlinearity in time series. *Biometrika* 73 (3), 687–694.
- Pfaff, B., 2008a. Analysis of Integrated and Cointegrated Time Series with R, second ed. Springer, New York. ISBN 0-387-27960-1. URL, <http://www.pfaffikus.de>.
- Pfaff, B., 2008b. Var, svar and svec models: implementation within R package vars. *J. Stat. Softw.* 27 (4), 1–32. URL, <http://www.jstatsoft.org/v27/i04/>.
- Phillips, P., 1986. Understanding spurious regressions in econometrics. *J. Econ.* 33 (3), 311–340. URL, <http://ideas.repec.org/a/eee/econom/v33y1986i3p311-340.html>.
- Phillips, P., Ouliaris, S., 1990. Asymptotic properties of residual based tests for cointegration. *Econometrica* 58 (1), 165–193.
- Samuelson, P.A., 1965. Proof that properly anticipated prices fluctuate randomly. *Ind. Manag. Rev.* 6, 41–49.
- Seo, M.H., 2006. Bootstrap testing for the null of no cointegration in a threshold vector error correction model. *J. Econ.* 127 (1), 129–150.

- Seo, M., 2008. Unit root test in a threshold autoregression: asymptotic theory and residual-based block bootstrap. *Economet. Theor.* 24 (06), 1699–1716. URL, http://ideas.repec.org/a/cup/theor/v24y2008i06p1699-1716_08.html.
- Seo, M.H., 2011. Estimation of nonlinear error correction models. *Economet. Theor.* 27 (2), 201–234. URL, <http://www.jstor.org/stable/27975474>.
- Seo, M., Linton, O., 2007. A smoothed least squares estimator for threshold regression models. *J. Econ.* 141 (2), 704–735. URL, <http://ideas.repec.org/a/eee/econom/v141y2007i2p704-735.html>.
- Sephton, P.S., 2003. Spatial market arbitrage and threshold cointegration. *Am. J. Agric. Econ.* 85 (4), 1041–1046. URL, <http://ideas.repec.org/a/bla/ajagec/v85y2003i4p1041-1046.html>.
- Sergio, L., GianCarlo, M., Gaetano, S.F., 2017. Threshold cointegration and spatial price transmission when expectations matter. *Agric. Econ.* 49 (1), 25–39. URL, <https://onlinelibrary.wiley.com/doi/abs/10.1111/agec.12393>.
- Sims, C.A., 1980. Macroeconomics and reality. *Econometrica* 48 (1), 1–48. URL, <http://ideas.repec.org/a/ecm/emetrp/v48y1980i1p1-48.html>.
- Stock, J.H., 1987. Asymptotic properties of least squares estimators of cointegrating vectors. *Econometrica* 55 (5), 1035–1056.
- Tong, H., 1978. On a threshold model. In: *Pattern Recognition and Signal Processing*. Sijhoff & Noordhoff, Amsterdam. Ch.
- Tong, H., 1990. *Non-Linear Time Series. A Dynamical System Approach*. Oxford Science Publications.
- Tong, H., 2011. Threshold models in time series analysis—30 years on. *Stat. Interface* 4, 107–118.
- Tong, H., 2015. Threshold models in time series analysis—some reflections. *J. Econ.* 189 (2), 485–491. (frontiers in Time Series and Financial Econometrics). URL, <http://www.sciencedirect.com/science/article/pii/S0304407615001177>.
- Tsay, R.S., 1989. Testing and modeling threshold autoregressive processes. *J. Am. Stat. Assoc.* 84 (405), 231–240. URL, <http://www.jstor.org/stable/2289868>.
- Von Cramon-Taubadel, S., 1998. Estimating asymmetric price transmission with the error correction representation: an application to the German pork market. *Eur. Rev. Agric. Econ.* 25, 1–18.
- Wang, M., Chan, N.H., Yau, C.Y., 2016. Nonlinear error correction model and multiple-threshold cointegration. *Stat. Sin.* 26, 1479–1498.
- Ward, R., 1982. Asymmetry in retail, wholesale and shipping point pricing for fresh vegetables. *Am. J. Agric. Econ.* 64 (2), 205–212.
- Zhu, H.-M., Li, S.-F., Yu, K., 2011. Crude oil shocks and stock markets: a panel threshold cointegration approach. *Energy Econ.* 33 (5), 987–994. URL, <http://www.sciencedirect.com/science/article/pii/S0140988311001320>.

Further reading

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL, <https://www.R-project.org/>.

Chapter 8

Econometric analysis of productivity: Theory and implementation in R

Robin C. Sickles^{a,*}, Wonho Song^b and Valentin Zelenyuk^c

^a*Department of Economics, Rice University, Houston, TX, United States*

^b*School of Economics, Chung-Ang University, Seoul, Republic of Korea*

^c*School of Economics, University of Queensland, Brisbane, QLD, Australia*

*Corresponding author: e-mail: rsickles@rice.edu

Abstract

Our chapter details a wide variety of approaches used in estimating productivity and efficiency based on methods developed to estimate frontier production using stochastic frontier analysis (SFA) and data envelopment analysis (DEA). The estimators utilize panel, single cross section, and time series data sets. The R programs include such approaches to estimate firm efficiency as the time-invariant fixed effects, correlated random effects, and uncorrelated random effects panel stochastic frontier estimators, time-varying fixed effects, correlated random effects, and uncorrelated random effects estimators, semiparametric efficient panel frontier estimators, factor models for cross-sectional and time-varying efficiency, bootstrapping methods to develop confidence intervals for index number-based productivity estimates and their decompositions, DEA and Free Disposable Hull estimators. The chapter provides the professional researcher, analyst, statistician, and regulator with the most up to date efficiency modeling methods in the easily accessible open source programming language R.

Keywords: Production (technical) efficiency, Stochastic frontier analysis, Data envelopment analysis, Panel data, Index numbers, Nonparametric analysis, Bootstrapping

1 Introduction

Our chapter provides a discussion of various statistical and mathematical procedures that firms, regulators, and academics and policy makers utilize in order to better understand performance and production (technical) efficiency of the entities they are benchmarking against other competitors or peers. As well we discuss, give examples of, and provide extensive links to R programs that

implement these various methods and approaches. The various methods and approaches distinguish themselves by leveraging the estimation of relative performance and of production efficiency measures on regression-based methods and on linear programming (LP) methods, the former referred to as stochastic frontier analysis (SFA) and the latter as data envelopment analysis (DEA). We discuss the main approaches in turn, their relative strengths and weaknesses, and briefly touch on ways to aggregate the various methods using model averaging approaches.

Our chapter is organized in the following way. We first briefly discuss the motivation for using such methods that rest on the presence of production efficiency differences among units of production that are being compared. The various sets of rationales for such a discipline as efficiency and productivity analysis can be categorized into three main groups of motivating factors. They are based on varying management practices, which deliver heterogeneous outcomes, behavioral economics, and the presence of X-efficiency. We then discuss the two main classes of production efficiency estimators: the SFA and the DEA estimators. We next discuss examples of R programs that implement these various classical estimators and various extensions that have been introduced in the recent literature as well as protocols for accessing R programs on well-documented websites and other open source sites. We then provide concluding remarks. Detailed technical discussions of many of the issues we discuss in our chapter can be found in [Sickles and Zelenyuk \(2019\)](#).

2 Why estimate production (technical) efficiency?

One of the most compelling rationales for the study of production efficiency is substantial heterogeneity in management practices and resulting changes in the operating efficiency of a firm. A consensus in the empirical literature exists that indicates substantial productivity differences both within a firm^a over time and among firms ([Foster et al., 2008](#); [Hall and Jones, 1999](#); [Hsieh and Klenow, 2009](#); [Lieberman et al., 1990](#)). [Glaister \(2014\)](#), among others, has noted that management practices are a key factor in explaining such productivity differences. Other factors, such as expenditures on R&D, utilization of capacity, and technology adoption, which are key decisions of management ([Nallari and Bayraktar, 2010](#)), are typically controlled by management practices. [Bloom et al. \(2012\)](#) regressed gross domestic product (GDP) per capita on a set of indicators of management practices among 17 countries. These indicators of management practices explained 87% of the variation in per capita GDP. These findings are corroborated in a more micro-oriented study of Indian firms by [Bloom et al. \(2013\)](#). The engineering mechanism, blueprints, formal structural statistical model, or economic model that explains

^aWe will use the term “firm” to denote any generic unit whose production efficiency is being measured and estimated.

how a firm's productivity is linked to management skills and practices has not been developed in a way that lends itself to empirical analysis and to the generation of relative technical efficiency differences among peer firms. We thus tend to view such a factor as an unobservable latent factor and assume that management practices are one of the key factors in firm productivity. Another factor is innovation but innovation is oftentimes facilitated by decisions and practices of management. The literature on the effects of management practices on production efficiency, labor efficiency, and related measures of a firm's financial success is quite dense.^b Behavioral economics provides another motivation for why firms may not operate at the frontier of production efficiency. The assumptions of efficient markets and rational decision makers have been leveraged with substantial success by neoclassical economists, but the footing on which these assumptions rest may be a bit loose and slippery.

X-efficiency theory is a pragmatic paradigm that admits to the prevalence of various sources on inefficiency in economics. It was introduced by Leibenstein (1966, 1975, 1987), who pointed out that agency problems, asymmetric information, and monitoring by regulators were all factors that generated incentives to engage in suboptimal decision-making, absent these factors and constraints. Drawing from studies of the health care industry, telecommunications, airlines, and education, Frantz (1997, 2007) has documented the ubiquitous existence of levels of inefficiency with the predictions from X-efficiency theory and has also concluded that such production inefficiency is much more significant than inefficiencies due to incorrect output and input allocations (allocative inefficiency). Other studies that have documented such inefficiency levels in the banking system can be found in Kwan (2006), Jiang et al. (2009), Fu and Heffernan (2009), Yao et al. (2008), Rezvanian et al. (2011), Bauer and Hancock (1993), Mester (1993), and DeYoung (1998). And as emphasized by Frantz (1997)

“...what becomes of the word maximize if non-maximize is not possible? Is the concept of efficiency important if the possibility of inefficiency is ruled out a priori? The importance of efficiency remains as long as economics remains important...”

X-efficiency theory and the methods that we present in our chapter is based on an interdisciplinary approach that combines psychology, management, statistics, applied mathematics, and engineering.

Finally, the usefulness and importance of SFA and DEA methods have passed the market test as they are required to be used in a wide variety of regulatory decision-making settings in Europe and elsewhere (Agrell et al., 2017; Bogetoft, 2013).

^bMuch of this literature is summarized and referenced in Grifell-Tatjé and Lovell (2013, 2015) and Grifell-Tatjé et al. (2018).

3 Regression-based methods to estimate production efficiency

We begin our formal discussion of production efficiency measurement by first defining a few sets and functions that are necessary for our presentation of methods and how to implement them. To facilitate further discussion, let inputs and outputs be represented by $x = (x_1, \dots, x_N)' \in \mathfrak{R}_+^N$ and $y = (y_1, \dots, y_M)' \in \mathfrak{R}_+^M$, respectively, and the production process characterized by a technology set T that satisfies standard regularity conditions (Sickles and Zelenyuk, 2019). The output set is defined as

$$P(x) \equiv \{y \in \mathfrak{R}_+^M : y \text{ is producible from } x \in \mathfrak{R}_+^N\}. \quad (1)$$

The production function $f : \mathfrak{R}_+^N \rightarrow \mathfrak{R}_+^1$ is defined as $f(x) \equiv \max\{y: y \in P(x)\}$. The maximum of the production function exists and is unique owing to the fact that $P(x)$ is a compact set. Details on the regularity conditions that assure the maximum is unique and exists also can be found in Sickles and Zelenyuk (2019). When there is more than one output one can use the generalization of the production function, *output-oriented* (Shephard, 1970) distance function $D_o : \mathfrak{R}_+^N \times \mathfrak{R}_+^M \rightarrow \mathfrak{R}_+^1 \cup \{+\infty\}$ as $D_o(x, y) \equiv \inf\{\theta > 0 : y/\theta \in P(x)\}$. For a technology with only one output and holding the inputs constant, D_o is the ratio of actual output to potential (maximal) output. Then frontier production is $f(x^o) = y^o/D_o(x^o, y^o)$. When there are multiple outputs, the Shephard output distance function is the smallest scalar required to radially expand *all* outputs to the output set's boundary, again at a fixed level of inputs.

3.1 The stochastic frontier paradigm

In cross-sectional SFA there are at least two sources of error (panel extensions may extend the identifiable error components for four distinct sources of error). In addition to the error that is appended to the parametric or nonparametric production or distance function to account for standard statistical noise (assumed to have mean zero) an additional source of error is added to address the asymmetry in the errors displayed by most production or distance function estimates that is due to the boundary property of the function being estimated. Such a one-sided error is utilized additively as a placeholder for production or technical inefficiency, which diminishes the level of observed output from the frontier production or distance function specified parametrically or nonparametrically. An example of such an SFA model for a linear in logs (Cobb–Douglas) production function is

$$\ln(y) = \beta_0 + \sum_{k=1}^N \beta_k \ln(x_k) + \varepsilon, \quad (2)$$

$$\varepsilon = v - u. \quad (3)$$

Here v is a random term with mean zero and variance σ_v^2 . The error term u represents the latent production inefficiency term. The latent production inefficiency term has positive support with a mean $\mu > 0$ and a variance σ_u^2 .^c Although these assumptions have been modified to account for more realistic empirical settings, this forms the basis for the canonical model used in SFA. Most of the literature in SFA has worked to generalize the distributions of u and v , their moment properties, stochastic properties of the inputs (and for multioutput distance functions the outputs as well), and different dependency patterns among the disturbances and the inputs and outputs, as well as possibly distinguishing between variables that are used in developing the production boundary and environmental or confounding factors that may influence production efficiency, that is, the deviations between the boundary and non-boundary observations, as opposed to influencing the boundary per se.

3.1.1 Corrected OLS

One of the first approaches to develop measures of production efficiency is the corrected ordinary least squares. The steps employed in such an exercise involve first estimating the production (or distance) function by ordinary least squares to get estimates of the average production relationship. In the second step one simply shifts the intercept (the example here is for the single-output production function) to ensure that the residuals are all nonpositive (Olson et al., 1980). With only mild assumptions, OLS will yield minimum variance unbiased linear estimators. The step with the Cobb–Douglas production function involves estimating

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_{i1}) + \dots + \beta_N \ln(x_{iN}) + \varepsilon_i, \quad i = 1, \dots, n. \quad (4)$$

The second step uses the intercept correction

$$\hat{\beta}_0^{cols} := \hat{\beta}_0 + \max_i \{\hat{\varepsilon}_i\}, \quad i = 1, \dots, n, \quad (5)$$

where $\hat{\varepsilon}_i$ are OLS residuals. The last step estimates production efficiency for a firm as

$$\text{Technical Inefficiency} = \hat{\varepsilon}_i^{cols} := \max_i (\hat{\varepsilon}_i) - \hat{\varepsilon}_i, \quad i = 1, \dots, n. \quad (6)$$

Thus the corrected production function that estimates the production frontier is

$$\ln(y_i) = \hat{\beta}_0^{cols} + \hat{\beta}_1 \ln(x_{i1}) + \dots + \hat{\beta}_N \ln(x_{iN}) - \hat{\varepsilon}_i^{cols}, \quad i = 1, \dots, n. \quad (7)$$

^cCost inefficiency can be modeled with a one-sided error with only negative support.

3.1.2 Stochastic frontier model

Aigner et al. (1977) (hereafter ALS) and Meeusen and van den Broeck (1977) pursued a parametric model of the stochastic frontier via maximum likelihood. Average inefficiency is defined in terms of the performance of a firm to the firm identified as having the best-practices, as measured by its level of efficiency. The original ALS model used a half-normal distribution for the efficiency term and a normal error for the idiosyncratic disturbance. Many other distributions have been considered, usually for the inefficiency term. These include the exponential, truncated normal, gamma, and doubly truncated normal (Almanidis and Sickles, 2012; Almanidis et al., 2014; Greene, 1980a, b; Qian and Sickles, 2008; Stevenson, 1980). In the canonical SFA the composite error terms are assumed to be independent. If the production function is linear in logs then a convenient parameterization is

$$y_i = f(x_i|\beta) \exp(\varepsilon_i), \quad i = 1, \dots, n \quad (8)$$

and for $\varepsilon = v - u$ the inefficiency of firm i is measured as

$$\exp(-u_i) \equiv \frac{y_i}{f(x_i) \exp(v_i)}, \quad i = 1, \dots, n. \quad (9)$$

The model is usually specified after log-transforming the production relationship

$$\ln y_i = \ln f(x_i|\beta) + v_i - u_i, \quad i = 1, \dots, n \quad (10)$$

and assuming

$$v_i \sim \mathcal{N}(0, \sigma_v^2)$$

and

$$u_i \sim |\mathcal{N}(0, \sigma_u^2)|$$

and that u_i and v_i are independent and i.i.d., it follows that the density of ε_i is:

$$f_{\varepsilon_i}(\varepsilon) = \frac{2}{\sigma} \phi\left(\frac{\varepsilon}{\sigma}\right) \left[1 - \Phi\left(\frac{\varepsilon\lambda}{\sigma}\right) \right], \quad -\infty \leq \varepsilon \leq +\infty \quad (11)$$

where $\Phi(\cdot)$ is the c.d.f. for the standard normal distribution function, $\sigma^2 = (\sigma_v^2 + \sigma_u^2)$, and $\lambda = \sigma_u/\sigma_v$. The composite errors, $\varepsilon_1, \dots, \varepsilon_n$, are of course not observed, but from our model we know that

$$\varepsilon_i = \ln y_i - \ln f(x_i), \quad i = 1, \dots, n. \quad (12)$$

The log-likelihood function is

$$\ell(y_1, \dots, y_n | \beta, \lambda, \sigma^2) = \frac{n}{2} \ln\left(\frac{2}{\pi}\right) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n [\ln y_i - \ln f(x_i)]^2 + \sum_{i=1}^n \ln \left[1 - \Phi\left(\frac{[\ln y_i - \ln f(x_i)]\lambda}{\sigma}\right) \right]. \quad (13)$$

Inefficiency has a mean and variance given by

$$E(u_i) \equiv \mu = \frac{\sqrt{2}}{\sqrt{\pi}} \sigma_u \quad (14)$$

and

$$V(u_i) = \left(\frac{\pi-2}{\pi}\right) \sigma_u^2 \quad (15)$$

and thus the composed error's mean and variance is

$$E(\varepsilon_i) = E(v_i - u_i) = E(-u_i) = -\mu = -\frac{\sqrt{2}}{\sqrt{\pi}} \sigma_u \quad (16)$$

$$V(\varepsilon_i) = V(v_i - u_i) = V(v) + V(u) = \sigma_v^2 + \left(\frac{\pi-2}{\pi}\right) \sigma_u^2 \quad (17)$$

and

$$\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j. \quad (18)$$

3.1.2.1 Estimation of individual inefficiencies

Cross-sectional SFA, unfortunately, does not provide a consistent estimate of the efficiency of each firm. A simple solution by [Materov \(1981\)](#) and further developed in [Jondrow et al. \(1982\)](#) uses $E(u_i | \varepsilon_i)$ as a point estimate of u_i . This estimator has its drawback as it is contaminated by statistical noise but it is appealing in that, from the law of iterated expectations,

$$E(u_i) = E_\varepsilon [E(u_i | \varepsilon_i)] \quad (19)$$

and thus to develop a consistent estimator of $E(u_i | \varepsilon_i)$ we can write

$$E(u_i | \varepsilon_i) \equiv \int_0^\infty u f_{u_i | \varepsilon_i}(u | \varepsilon) du \quad (20)$$

where, from the definition of the conditional density,

$$f_{u_i | \varepsilon_i}(u | \varepsilon) \equiv \frac{f_{u_i, \varepsilon_i}(u, \varepsilon)}{f_{\varepsilon_i}(\varepsilon)}. \quad (21)$$

Based on the normal/half-normal composed error structure typically used in SFA we can write

$$E(u_i|\varepsilon_i) = \mu_* + \sigma_* \frac{1}{1 - \Phi(-\mu_*/\sigma_*)} \phi\left(\frac{-\mu_*}{\sigma_*}\right) \quad (22)$$

or

$$E(u_i|\varepsilon_i) = \sigma_* \left[\frac{\mu_*}{\sigma_*} + \frac{1}{1 - \Phi(-\mu_*/\sigma_*)} \phi\left(\frac{-\mu_*}{\sigma_*}\right) \right]. \quad (23)$$

Using the parameterization $-\frac{\mu_*}{\sigma_*} = \frac{\sigma_v^2 \varepsilon}{\sigma_v^2 \sigma_u} = \frac{\sigma_u \varepsilon}{\sigma_v} = \frac{\varepsilon \lambda}{\sigma}$, the conditional expectation is:

$$E(u_i|\varepsilon_i) = \frac{\sigma_v \sigma_u}{\sigma} \left[-\frac{\varepsilon_i \lambda}{\sigma} + \frac{\phi(\varepsilon_i \lambda / \sigma)}{1 - \Phi(\varepsilon_i \lambda / \sigma)} \right] \quad (24)$$

for which a consistent estimate can be obtained based on consistent estimates of the model parameters (e.g., obtained via MLE or COLS, or the Modified OLS procedures of [Olson et al. \(1980\)](#)). Thus a (conditional) consistent estimator of each firm's inefficiency level can be based on this last equation.

3.1.3 Panel stochastic frontiers

The cross-sectional stochastic frontier model has a number of drawbacks that have been addressed over the 40 years since it was introduced. As we pointed out, a consistent estimate for a firm's technical efficiency is not available, only a consistent estimate of the conditional mean of the firm's efficiency level. Of course the canonical ALS model is fully parametric and inputs are assumed to be uncorrelated with the regressors, which is particularly troubling when the regressors are inputs and the latent factor that typically is assumed to account for inefficiency is unobservable managerial expertise, an input whose independence from the levels of capital and labor used in production is a questionable assumption, and of course if a firm is aware of its level of technical efficiency this information, although unknown to the analyst, should not be independent of the firm's input choices.

These problems are potentially avoidable if one has panel data ([Pitt and Lee, 1981](#); [Schmidt and Sickles, 1984](#)), although the possible endogeneity of input choice must be addressed with care and its treatment using fixed effects type estimators may not be completely satisfactory. We will return to this issue shortly. If we have a panel of firms, then if the unobserved firm effects represent technical efficiency they can be estimated consistently for large T (assuming that our sample of firms n is large) after controlling for inputs and other environmental and observable factors that may impact production. To show how this is accomplished with panel data we use the general treatment discussed in [Schmidt and Sickles \(1984\)](#) (SS), which also considers the [Pitt and Lee \(1981\)](#) parametric random effects model. The model is

$$y_{it} = \alpha + x'_{it} \beta + v_{it} - u_i, \quad i = 1, \dots, n; \quad t = 1, \dots, T, \quad (25)$$

which can be rewritten as

$$y_{it} = \alpha^* + x'_{it}\beta + v_{it} - u_i^* \quad (26)$$

where $\alpha^* = \alpha - \mu$; $u_i^* = u_i - \mu$; $E(u_i) = \mu \geq 0$, and where x_{it} is a vector of N inputs. If we let $\alpha_i = \alpha^* - u_i^*$ then the model becomes the usual panel data model

$$y_{it} = \alpha_i + x'_{it}\beta + v_{it} \quad (27)$$

and cross-sectional effects that can be viewed as random, fixed, or simply ignored. Five estimators of the classical panel data model with time-invariant effects are discussed in SS. These are the pooled OLS model, the fixed effects within estimator, the random effects model, the Hausman–Taylor estimator, and the fully parametric random effects MLE model (Pitt and Lee, 1981). Schmidt and Sickles (1984) discuss the asymptotics of each of these estimators and the assumptions needed in order for the parameter estimates and estimates of the technical efficiency level to be consistently estimated.

Technical efficiency effects estimates with the SS estimators do not change over time and this is a strong and often unreasonable assumption that is not necessary. The suite of panel stochastic frontier estimators developed by Cornwell et al. (1990) (CSS) had a parameterization that allowed for time-varying heterogeneity and is based on the model:

$$y_{it} = x'_{it}\beta + z'_i\gamma + w'_{it}\delta_i + v_{it}, \quad i = 1, \dots, n; \quad t = 1, \dots, T, \quad (28)$$

where x_{it} , z_i , and w_{it} are $N \times 1$, $J \times 1$, and $L \times 1$ vectors, respectively, and the parameter vectors β , γ , and δ_i are dimensioned conformably. With $\delta_0 = E[\delta_i]$, and $\delta_i = \delta_0 + u_i$ the model can be written as:

$$y_{it} = x'_{it}\beta + z'_i\gamma + w'_{it}\delta_0 + \varepsilon_{it},$$

where

$$\varepsilon_{it} = v_{it} + w'_{it}u_i \quad (29)$$

and where u_i is assumed to be *i.i.d.*, zero mean random variables with covariance matrix Δ . The error term v_{it} is assumed to be *i.i.d.*, with zero mean and constant variance σ_v^2 . The basic model assumes that v_{it} is uncorrelated with z , x , and u_i . In order to allow for time-varying technical efficiency the stochastic frontier model of Schmidt and Sickles (1984)

$$y_{it} = \alpha + x'_{it}\beta + v_{it} - u_i = \alpha_i + x'_{it}\beta + v_{it} \quad (30)$$

can be modified by replacing the α_i with, e.g.,

$$\alpha_{it} = \theta_{i1} + \theta_{i2}t + \theta_{i3}t^2, \quad (31)$$

and the model in matrix form becomes

$$y = X\beta + Z\gamma + W\delta_0 + \varepsilon, \quad (32)$$

$$\varepsilon = Qu + v. \quad (33)$$

Details of the estimator can be found in [Cornwell et al. \(1990\)](#). Fixed effects, random effects, and Hausman–Taylor estimators of the CSS model were developed to estimate productivity efficiency that is time varying and allow for consistent estimation under large n and T asymptotics of the time-varying productivity efficiency for each firm. The parameter δ_i is estimated by regressing the residuals $(y_{it} - x_{it}'\beta)$ for firm i on w_{it} . This amounts to regressing the within residual on a constant term, time, and time-squared for the specification we introduced for α_{it} above and relative inefficiencies can be approximated (in the linear in logs production function) by

$$\hat{\alpha}_t = \max_i (\hat{\alpha}_{it}), \quad i = 1, \dots, n \quad (34)$$

$$\hat{u}_{it} = \hat{\alpha}_t - \hat{\alpha}_{it}. \quad (35)$$

Parametric MLE can also be used to estimate random effects time-varying technical efficiency models. [Kumbhakar \(1990\)](#) and [Battese and Coelli \(1992\)](#) present two such models. In the former, inefficiency is modeled as

$$u_{it} = (1 + \exp(bt + ct^2))^{-1} \tau_i, \quad (36)$$

where a and b are parameters to be estimated and where τ_i 's distribution is assumed to be $i.i.d N^+(0, \sigma_\tau^2)$ and v_{it} is $i.i.d \mathcal{N}(0, \sigma_v^2)$. In the latter specification inefficiency is modeled as

$$u_{it} = \eta_i \tau_i = \{\exp[-\eta(t-T)]\} \tau_i, \quad i = 1, \dots, n; \quad t = 1, \dots, T, \quad (37)$$

where v_{it} is $i.i.d \mathcal{N}(0, \sigma_v^2)$ random variables, τ_i is assumed to be $i.i.d$ and has a nonnegative truncated distribution $\mathcal{N}(\mu, \sigma^2)$, and η is a scalar parameter.

3.1.4 Second- and third-generation stochastic frontier models

The models we have just discussed provide the basic statistical intuition for SFA. Much has been done to extend and generalize the canonical models we have discussed and space prohibits us from providing comparable detail on these second- and third-generation extensions. However, the R codes that we will discuss shortly are equipped to deal with a number of these relatively new modeling scenarios and so we give a brief summary of what these second- and third-generation SFA models deliver and the general ideas behind how they are formulated and specified. We leave it to the reader to seek out the original sources and recent books, handbooks, and survey articles that we have referenced and that provide more detailed treatments.

Researchers have provided further generalizations of the CSS time-varying technical efficiency model that focus on two main aspects of the model. The first is to allow for a factor-type structure for the cross section and time-varying technical efficiency term (these are almost all exclusively panel data extensions). The second is to decompose the time-invariant and time-varying efficiency terms into a portion that is skewed, to represent technical

inefficiency, and a portion that is symmetric, the latter somewhat courageously called the “true” effects model. There are four other generalizations that some may view as comparable to these but there is simply not enough space to address them in any substantive way. We will mention them and provide several references as well as links to R codes that deal with these four additional issues, which are environmental factors, endogeneity, Bayesian methods for SFA, and nonparametric specifications of the technology and of the error structure.

We briefly discuss these sets of generalizations and available R code to implement them.

3.1.4.1 Factor models and SFA

[Lee \(1991\)](#) and [Lee and Schmidt \(1993\)](#) were the first to propose a (one component) factor model to address time-varying and cross-sectional specific production efficiency. Their model is

$$y_{it} = \alpha_t + x'_{it}\beta + v_{it} - u_{it} \quad \text{for } i = 1, \dots, n; t = 1, \dots, T, \quad (38)$$

or as

$$y_{it} = x'_{it}\beta + \alpha_{it} + v_{it}, \quad (39)$$

where $\alpha_{it} = \alpha_t - u_{it}$ is the time-varying cross-sectional specific technical efficiency term, which are modeled as

$$\alpha_{it} = \eta_t \delta_i. \quad (40)$$

Here η_t , $t = 1, \dots, T$, are the time-varying effects to be estimated and δ_i , $i = 1, \dots, n$, are the firm effects. [Lee and Schmidt \(1993\)](#) provide fixed effects and random effects estimators for this one-factor model based on nonlinear regression estimators.

[Ahn et al. \(2007\)](#) generalized this model to allow for multiple factors that change over time and specify the production frontier as

$$y_{it} = \delta_t + x'_{it}\beta + v_{it} - u_{it} \quad (41)$$

$$= x'_{it}\beta + \eta_{it} + v_{it}, \quad i = 1, 2, \dots, n, \quad t = 1, 2, \dots, T, \quad (42)$$

where, again, v_{it} is the usual disturbance term and $u_{it} \geq 0$ is the inefficiency term and where $\eta_{it} \equiv \delta_t - u_{it}$ is the time-varying and cross-sectional specific technical efficiency term that is expressed as a linear combination of p unrestricted components,

$$\eta_{it} = \theta_{1t}\alpha_{1i} + \theta_{2t}\alpha_{2i} + \dots + \theta_{pt}\alpha_{pi} = \sum_{j=1}^p \theta_{ji}\alpha_{ji}. \quad (43)$$

The model is estimated using generalized methods of moments. [Ahn et al. \(2013\)](#) provide a focused study of consistency properties of their [Ahn et al. \(2007\)](#)

model when different sorts of dependency relationships exist between the production efficiency effects and the regressors.

In the Ahn et al. (2007) model the effects are multiplicative. Kneip et al. (2012) (KSS) provided a more general model than Ahn et al. (2007) by allowing for a general nonparametric time-varying and cross section specific productivity efficiency. The KSS model is

$$y_{it} = \beta_0(t) + \sum_{j=1}^N \beta_j x_{ij} + u_i(t) + v_{it}, \quad i = 1, \dots, n; \quad t = 1, \dots, T. \quad (44)$$

The $u_i(t)$'s are assumed to be smooth time-varying individual effects that satisfy a normalization that $\sum_i u_i(t) = 0$ and are a linear combination of $L < T$ basis functions (common factors) g_1, \dots, g_L :

$$u_i(t) = \sum_{r=1}^L \theta_{ir} g_r(t). \quad (45)$$

The term $\beta_0(t)$ is an average function that is eliminated by centering the model yielding

$$y_{it} - \bar{y}_t = \sum_{j=1}^N \beta_j (x_{ij} - \bar{x}_{ij}) + u_i(t) + v_{it} - \bar{v}_i, \quad i = 1, \dots, n; \quad t = 1, \dots, T \quad (46)$$

where $\bar{y}_t = \frac{1}{n} \sum_i y_{it}$, $\bar{x}_{ij} = \frac{1}{n} \sum_i x_{ij}$, and $\bar{v}_i = \frac{1}{T} \sum_t v_{it}$. As with the CSS and many other panel stochastic frontier estimators, technical efficiency is calculated as $TE_i(t) = \exp \{u_i(t) - \max_{j=1, \dots, n} (u_j(t))\}$, just as it is calculated with the CSS estimator. The estimator used is a three-step one that first estimates β using penalized least squares and smoothing splines to approximate the factors, then second generates estimates of the covariance structure of the $u_i(t)$'s, and the third step estimates the basis functions \hat{g}_r and then updates estimates of u_i by $\sum_{r=1}^L \hat{\theta}_{ir} \hat{g}_r$. The KSS estimator and related estimators such as those introduced by Bai and Ng (2002) and Bai (2009), as well as tests for the number of factors, have been programmed in R and are available not only in the suite of programs we discuss at the end of this chapter but also can be found on the website referenced in Bada and Liebl (2014).

3.1.4.2 True fixed effects and SFA

Greene (2005a, b) proposed a stochastic panel frontier model in which the intercept fixed effects were not measures of persistent inefficiencies, as had been assumed by SS, but rather was simply firm-specific heterogeneity. This has become known as the “true” fixed effects model, although a less ambitious label may be the panel frontier with only transitory inefficiency. The model is

$$y_{it} = \alpha_i + x'_{it} \beta + v_{it} - u_{it} \quad (47)$$

$$\varepsilon_{it} = v_{it} - u_{it}, \quad (48)$$

or

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it} \quad (49)$$

and thus the source of inefficiency is u_{it} , which is distinguished from the intercept terms α_i . Greene outlines estimators for both the fixed effects and random effects panel stochastic frontier model with only transitory inefficiency based on simulated MLE and technical efficiency estimates are based on the familiar expression

$$\widehat{TE}_{it} = \exp[-\{\max(\hat{u}_{it}) - \hat{u}_{it}\}]. \quad (50)$$

[Colombi et al. \(2011, 2014\)](#) and [Tsionas and Kumbhakar \(2014\)](#) provided further generalizations of this model and alternative estimators based on Bayesian methods, full information maximum likelihood, quasi-MLE, and method of moments. One important generalization proposed by these authors specifies the persistent and transitory efficiency terms as separate skew-normal errors based on the following parameterization of [Colombi et al. \(2014\)](#)

$$y_{it} = x'_{it}\beta + \alpha_i + v_{it} - u_{it} - \eta_i, \quad (51)$$

where $\varepsilon_{it} = \alpha_i + v_{it} - u_{it} - \eta_i$ is the error structure and where u_{it} and η_i are non-negative random variables that capture firm-specific random effects, random noise, short-run technical inefficiency, and long-run technical inefficiency.

Environmental factors can be introduced into the SFA paradigm in a natural way by including them in a vector of variables that is assumed to impact the mean and the variance of production efficiency. For an excellent treatment of the problem see [Wang and Schmidt \(2002\)](#), [Simar and Wilson \(2008\)](#), and [Kim and Schmidt \(2008\)](#). The marginal effects of the environmental factors, based on [Battese and Coelli \(1995\)](#), can be estimated using the R package Frontier (<http://cran.r-project.org/web/packages/frontier/>) developed by Arne Henningsen.

Various dependency structures have been assessed within the SFA paradigm. An excellent source of recent work on this topic can be found in [Kumbhakar and Schmidt \(2016\)](#).

No one has had more of an impact on the introduction of Bayesian methods into the SFA and DEA paradigms than Mark Steel and Mike Tsionas ([Griffin and Steel, 2007](#); [Liu et al., 2017](#); [Tsionas and Papadakis, 2010](#)). The WinBUGS package of Steele is Fortran and MATLAB®-based open source code that can be modified to run in R. Nonparametric models in SFA have been studied in a variety of settings by [Park et al. \(1998, 2003, 2007\)](#), [Adams et al. \(1999\)](#), among others, and the R codes for the latter models are discussed in more detail in the last sections of this chapter on specific implementation of open source R codes for SFA and DEA. Experimental

R codes that can estimate both parametric and nonparametric stochastic frontiers can also be found in the experimental R-Forge package maintained by Arne Henningsen.

4 Envelopment estimators

In this section we will briefly review another very popular approach in the measurement and empirical estimation of the efficiency of economic systems (firms, industries, etc.), called DEA.

4.1 The origins of DEA

In a nutshell, DEA is rooted in and coherent with theoretical economic modeling using *activity analysis models* (AAM) and is estimated via the powerful LP approach. Its name was branded in the seminal work of Charnes et al. (1978), who refined and generalized the approach of Farrell (1957) to estimate production efficiency, which in turn was influenced by seminal works of Debreu (1951), Koopmans (1951a, b), and Shephard (1953, 1970).

4.2 The basic DEA model

In his seminal work, Farrell (1957) focused on the constant returns to scale model, with multiple inputs and a single output. About two decades later, Charnes et al. (1978) generalized Farrell's approach to the multioutput case but started with a very different formulation—a fractional programming problem formulation with the objective being to optimize the ratio of a weighted aggregate of outputs to a weighted aggregate of inputs (i.e., a kind of productivity index). They then transformed this problem into an LP problem and derived its dual, which turned out to be the (generalized version of) AAM proposed by Farrell. Specifically, the formulation of Charnes et al. (1978) for estimating the efficiency score of a firm or decision-making unit (DMU) $j \in (1, \dots, n)$ with an allocation (x^j, y^j) , states

$$E_{i,CCR}^j = \max_{v_1, \dots, v_N; u_1, \dots, u_M} \left\{ \frac{\sum_{m=1}^M u_m y_m^j}{\sum_{l=1}^N v_l x_l^j} : \frac{\sum_{m=1}^M u_m y_m^k}{\sum_{l=1}^N v_l x_l^k} \leq 1, \quad k = 1, \dots, n, \right. \\ \left. u_m \geq 0, v_l \geq 0, \quad l = 1, \dots, N; \quad m = 1, \dots, M \right\} \quad (52)$$

where $u' = (u_1, \dots, u_M)$ and $v' = (v_1, \dots, v_N)$ are optimization variables (also called here “multipliers”). This DEA formulation is more popular in the operations research and management science literature and is often referred to as the “CCR model” or the “multiplier form of DEA” under CRS, additivity and free disposability.

Importantly, Charnes et al. (1978) then showed that (52) is equivalent to the AAM version of the DEA-estimator of the Farrell input oriented technical efficiency score of a DMU with an allocation (x^j, y^j) under the assumption of CRS (also assuming additivity and free disposability of all inputs and all outputs), formulated as

$$\hat{E}_i(x^j, y^j) \equiv \min_{\lambda, z^1, \dots, z^n} \lambda \quad (53)$$

s.t.

$$\sum_{k=1}^n z^k y_m^k \geq y_m^j, \quad m = 1, \dots, M, \quad (54)$$

$$\sum_{k=1}^n z^k x_l^k \leq \lambda x_l^j, \quad l = 1, \dots, N, \quad (55)$$

$$\lambda \geq 0, z^k \geq 0, \quad k = 1, \dots, n, \quad (56)$$

which is an LP problem that can be solved via any standard LP solver. This formulation is more common in the economics literature (largely due to its connection to works of Debreu (1951) and Koopmans (1951a, b)), and is often referred to as the *envelopment form of DEA* under CRS, additivity and free disposability. We will use this envelopment formulation to describe other variants of DEA, though it is useful to keep in mind that there is also a multiplier form that optimizes what can be viewed as a normalized productivity index.

The previous formulation looks at minimization of all inputs, while keeping outputs fixed, and hence is called the input orientation. Similarly, the DEA-estimator of the Farrell output-oriented technical efficiency score of any (x^j, y^j) allocation, under the assumptions of CRS, and additivity and free disposability of all outputs and all inputs is formulated as

$$\hat{E}_o(x^j, y^j) \equiv \max_{\lambda, z^1, \dots, z^n} \lambda \quad (57)$$

s.t.

$$\sum_{k=1}^n z^k y_m^k \geq \lambda y_m^j, \quad m = 1, \dots, M, \quad (58)$$

$$\sum_{k=1}^n z^k x_l^k \leq x_l^j, \quad l = 1, \dots, N, \quad (59)$$

$$\lambda \geq 0, z^k \geq 0, \quad k = 1, \dots, n. \quad (60)$$

Immediately note that for these formulations we have

$\hat{E}_i(x^j, y^j) = 1/\hat{E}_o(x^j, y^j)$ for any (x^j, y^j) , as is required theoretically due to CRS.

Also note that the reciprocals of these estimated Farrell efficiency measures also give estimates of the input and output oriented Shephard's distance functions (Shephard, 1953, 1970), under the same assumptions on technology, i.e., CRS, additivity and free disposability.

Sometimes a researcher may not be interested in fixing only the levels of inputs or outputs, but may want to simultaneously expand outputs and contract inputs, thus requiring other orientations. The DEA estimator of a general efficiency measure of any (x^j, y^j) allocation in such a case (also here under the assumptions of CRS and additivity and free disposability of all outputs and all inputs) can be obtained as follows

$$\widehat{GE}_o(x^j, y^j) \equiv \max_{\lambda_1, \dots, \lambda_N, \theta_1, \dots, \theta_M, z^1, \dots, z^n} f(\lambda_1, \dots, \lambda_N, \theta_1, \dots, \theta_M) \quad (61)$$

s.t.

$$\sum_{k=1}^n z^k y_m^k \geq \theta_m y_m^j, \quad m = 1, \dots, M, \quad (62)$$

$$\sum_{k=1}^n z^k x_l^k \leq x_l^j / \lambda_l, \quad l = 1, \dots, N, \quad (63)$$

$$\lambda \geq 0, z^k \geq 0, \quad k = 1, \dots, n, \quad (64)$$

where certain restrictions can be imposed on the objective function $f(\lambda_1, \dots, \lambda_N, \theta_1, \dots, \theta_M)$ and its arguments, depending on the interest of the researcher. For example, restricting f to be additive (but summing only positive arguments) will give the DEA estimate of the general Russell efficiency measure and an additional restriction of $\theta_1 = \dots = \theta_M = 1$ will turn it into the input-oriented Russell efficiency measure (as was originally introduced by Färe and Lovell (1978)). If one instead restricts $\lambda_1 = \dots = \lambda_N = 1$ then one obtains the output-oriented Russell efficiency measure. Meanwhile, restricting $f(\lambda_1, \dots, \lambda_N, \theta_1, \dots, \theta_M)$ to be multiplicative (a geometric mean) will generate the multiplicative-Russell efficiency measure introduced by Färe et al. (2007). Furthermore, if instead one imposes $\theta_1 = \dots = \theta_M = \theta$ and $\lambda_1 = \dots = \lambda_N = \lambda$ then the DEA estimate of the what has been termed the general hyperbolic efficiency measure is obtained. This latter measure sometimes also appears with the additional restriction that $\theta = \lambda$, which imposes the properties of equiproportional expansion (contraction) of all output (inputs). Note that in general, this last formulation is no longer an LP problem and therefore, it is typically more challenging to estimate.

Another very general measure, as well as a primal characterization of technology, is the directional distance function,^d which can be estimated via the

^dThe origins of ideas for this function go back to at least Allais (1943), Diewert (1983), and Luenberger (1992) and were later revived and thoroughly developed by Chambers et al. (1996).

following DEA formulation (here also under the assumptions of CRS and additivity and free disposability of all outputs and all inputs):

$$\widehat{D}_d(x^j, y^j | d_x, d_y) \equiv \max_{\lambda, z^1, \dots, z^n} \lambda, \quad (65)$$

s.t.

$$\sum_{k=1}^n z^k y_m^k \geq y_m^j + \lambda d_{y_m}, \quad m = 1, \dots, M, \quad (66)$$

$$\sum_{k=1}^n z^k x_l^k \leq x_l^j - \lambda d_{x_l}, \quad l = 1, \dots, N, \quad (67)$$

$$\lambda \geq 0, z_k \geq 0, \quad k = 1, \dots, n. \quad (68)$$

In a similar fashion, DEA can be used to model and estimate cost, revenue and profit functions and associated efficiency measures. More details on this can be found in [Sickles and Zelenyuk \(2019\)](#).

4.3 The myriad of DEA models

The approach briefly described in the previous section in the context of different efficiency measures constitutes the canonical forms of the DEA paradigm. Essentially, all the other versions are modifications or extensions of the models we have outlined. In this section we briefly discuss a few of these modifications and extensions.

Oftentimes, extensions are obtained by imposing various additional constraints onto either the envelopment form or the multiplier form of DEA, with an aim to better mimic the particular actual production process under study. Such additional restrictions should be imposed with care, since they may (and often do) affect other desirable properties related to previously added constraints, such as CRS, convexity, free disposability, additivity, etc.

Additional constraints also may create computational complications, e.g., turning the problem from a linear to a nonlinear one or a hybrid problem that may require integer-programming problems. This may result in possibly making the problem much harder to compute, possibly infeasible, or may give inferior local optima or degenerate solutions.

4.3.1 Relaxing constant returns to scale and convexity

The first wave of extensions of the DEA-CRS model mainly focused on relaxing assumptions such as CRS and convexity. This line of research was pursued by a number of researchers who extended and enriched the DEA paradigm, among them [Afriat \(1972\)](#), [Färe et al. \(1983\)](#), [Banker et al. \(1984\)](#), [Deprins et al. \(1984\)](#), [Petersen \(1990\)](#), and [Bogetoft \(1996\)](#), to mention a few.

Out of the many modifications and extensions of the canonical DEA problem, the several that have sustained the test of time and popularity are the DEA-VRS (variable returns to scale) and the DEA-NIRS (nonincreasing returns to scale) models, which simply amount to adding additional constraints in the form of $\sum_{k=1}^n z_k = 1$ or $\sum_{k=1}^n z_k \leq 1$, respectively, to the DEA-CRS formulations as described above.

Meanwhile, the free disposal hull (FDH) approach can be implemented via a hybrid of the LP and the integer-programming problems, which is formulated in exactly the same way as the DEA-VRS programming problem with the exception that the constraints “ $z_k \geq 0, k = 1, \dots, n$ ” are replaced with “ $z_k \in \{0, 1\}, k = 1, \dots, n$ ”. For example, using the output-oriented Farrell efficiency with an allocation (x^j, y^j) the FDH formulation is given by

$$\widehat{E}_o(x^j, y^j) \equiv \max_{\theta} \theta \quad (69)$$

s.t.

$$\sum_{k=1}^n z^k y_m^k \geq \theta y_m^j, \quad m = 1, \dots, M, \quad (70)$$

$$\sum_{k=1}^n z^k x_l^k \leq x_l^j, \quad l = 1, \dots, N, \quad (71)$$

$$\sum_{k=1}^n z^k = 1, \quad (72)$$

$$\theta \geq 0, z_k \in \{0, 1\}, \quad k = 1, \dots, n. \quad (73)$$

The value of such formulation is that it hints at the relationship between FDH and DEA-VRS: indeed, the DEA-VRS estimated technology set is simply the *convex closure* of the FDH-estimated technology set. Thus, the FDH estimator can be viewed as a special case of the data envelopment analysis approach since it also envelopes the data but does so without imposing convexity. For historical reasons, these names are kept separate to avoid confusion. An alternative yet equivalent form of the FDH estimator can be given via the min–max problem, which is faster to compute (e.g., see [Simar and Wilson \(2013\)](#) for more details).

Finally, stochastic versions of DEA and FDH are also available (e.g., [Simar, 2007](#), [Simar and Zelenyuk, 2011](#)).

4.3.2 Modeling with undesirable outputs or with congesting inputs

Another important stream of DEA literature has focused on estimating technologies with weak disposability of inputs or (and especially) outputs, to account for the fact that some outputs are undesirable (bad) and some inputs can cause congestion.

The ideas for such modeling approaches go back to at least [Shephard \(1974\)](#), and then was elaborated on in [Färe and Svensson \(1980\)](#), [Färe and Grosskopf \(1983\)](#), [Grosskopf \(1986\)](#), [Tyteca \(1996\)](#), and [Chung et al. \(1997\)](#), which defined the mainstream approach on this matter. More recently this mainstream approach was reevaluated in several important works, including [Seiford and Zhu \(2002\)](#), [Färe and Grosskopf \(2003\)](#), [Färe and Grosskopf \(2004\)](#), [Färe and Grosskopf \(2009\)](#), [Førsund \(2009\)](#), [Podinovski and Kuosmanen \(2011\)](#), [Pham and Zelenyuk \(2018\)](#), and [Pham and Zelenyuk \(2019\)](#).^e

While many proposals have been made, the most popular approach in this context so far continues to be the mainstream one. In this approach, for example, if one is interested in measuring the radial expansion of the good outputs (g) while having no more inputs (x) and no more of bad outputs (b), then the DEA estimate of the good-output-oriented Farrell technical efficiency (under VRS) is given by

$$\widehat{TE}_g(x, g, b) \equiv \max_{\gamma, z^1, \dots, z^n, \delta} \gamma \quad (74)$$

$$x \geq \sum_{k=1}^n z^k x^k \quad (75)$$

$$g\gamma \leq \delta \sum_{k=1}^n z^k g^k, \quad \gamma \geq 1 \quad (76)$$

$$b = \delta \sum_{k=1}^n z^k b^k, \quad 0 \leq \delta \leq 1 \quad (77)$$

$$\sum_{k=1}^n z^k = 1, \quad z^k \geq 0, \quad k = 1, \dots, n. \quad (78)$$

For more theoretical and practical (computational) discussions on this topic see [Pham and Zelenyuk \(2019\)](#) and references therein.

4.3.3 Other streams of DEA

Another stream of DEA focuses on accounting for the network structure of production technologies whether static or dynamic. This stream originated in the seminal works of [Färe and Grosskopf \(1996\)](#) and [Färe et al. \(1996\)](#) and was taken further in many other works, e.g., see [Kao \(2009a, b, 2014\)](#) and references therein.

Another important stream of DEA literature is on the topic of weight restrictions in multiplier form of DEA and the classical works here are by [Dyson and Thanassoulis \(1988\)](#), [Charnes et al. \(1990\)](#), and [Thompson et al. \(1990\)](#), with more recent and fundamental contributions including new interpretations

^eAlso see [Dakpo et al. \(2017\)](#) and [Sueyoshi et al. \(2017\)](#) for reviews of this research stream.

(as technological trade-offs) of various weight restrictions in DEA from [Podinovski and Bouzdine-Chameeva \(2013\)](#), to mention just a few. Also see reviews on this topic by [Allen et al. \(1997\)](#) and [Podinovski \(2015\)](#).

Yet another interesting research stream of DEA overlaps with game theory, which also has its roots in the seminal work of [von Neumann \(1945\)](#) and more explicit treatment, for example, in [Hao et al. \(2000\)](#), [Nakabayashi and Tone \(2006\)](#), [Liang et al. \(2008\)](#), and [Lozano \(2012\)](#).

4.4 Statistical analysis of DEA and FDH

Another key research wave that brought DEA and FDH to a totally different level is related to their statistical aspects—this wave was mainly influenced by the seminal works of Léopold Simar and many of his coauthors.

The first breakthrough in this area was made by [Banker \(1993\)](#), where the first proof of consistency of the DEA estimator was sketched for the single output case (in output oriented context), and was pointed out that it belongs to the class of maximum likelihood estimators. This important discovery was then substantially enriched by [Korostelev et al. \(1995a, b\)](#) who proved convergence of the estimated technology to the true technology for both DEA and FDH estimators, and derived convergence rates of these estimators, clarifying that they depend on the dimension of the production model, yet also have some optimality properties under certain conditions.

The convergence properties for the multiinput–multioutput case were first presented in the seminal work of [Kneip et al. \(1998\)](#). Meanwhile, the discovery of the limiting distribution of the DEA estimator was done by [Gijbels et al. \(1999\)](#), only for the single input/single output case and a decade later, [Kneip et al. \(2008\)](#) derived it for the fully multivariate case for DEA with VRS and also proved consistency of various bootstrap procedures. The limiting distribution for the case of DEA with CRS was established by [Park et al. \(2010\)](#) and the limiting distribution of the FDH estimator for the fully multivariate case was also derived by [Park et al. \(2000\)](#), while consistency of the bootstrap for FDH estimator was first presented in [Jeong and Simar \(2006\)](#).

This stream also includes the approach of analyzing the DEA (or FDH) estimated efficiency scores. Perhaps the most popular of these is the so-called two-stage DEA, which involves regression analysis of efficiency scores on some factors. The state of the art here is the approach proposed by [Simar and Wilson \(2007\)](#), which is based on truncated regression where the inference is done with the help of a double bootstrap.^f A nonparametric version of this approach (based on nonparametric truncated regression) was proposed by [Park et al. \(2008\)](#). Furthermore, methods to analyze the distributions of DEA and FDH efficiency scores were explored in [Simar and](#)

^fAlso see [Simar and Wilson \(2011\)](#) for the discussion on caveats and limitations.

Zelenyuk (2006), while methods to analyze industry efficiency were explored by Simar and Zelenyuk (2007). These approaches were applied in various contexts and industries, e.g., Zelenyuk and Zheka (2006), Demchuk and Zelenyuk (2009), Curi et al. (2015), Chowdhury and Zelenyuk (2016), and Du et al. (2018), to mention a few.

A particularly notorious drawback of DEA and FDH—not allowing for noise and sensitivity to “super-efficient” outliers—was addressed by Simar (2007) and Simar and Zelenyuk (2011), who proposed their version of Stochastic DEA and Stochastic FDH, which consists of two stages: (i) filter the data from the noise using a nonparametric stochastic frontier method^g and then (ii) use DEA or FDH on the filtered data.

More recently, Kneip et al. (2015) derived new central limit theorems for the context where DEA or FDH estimates are used in place of the true efficiency and thus provided the foundation for many useful statistical tests involving DEA or FDH estimators, including the two-stage “DEA +regression” context. This foundation was then used by Kneip et al. (2016) and Daraio et al. (2018b) to develop various statistical tests and by Simar and Zelenyuk (2018) to develop two new central limit theorems for the aggregate efficiency scores (industry efficiency, etc.) of the type described above and more work continues in this area.

5 SFA efficiency software in R

Reproducing the R code needed to implement the methods we have discussed in our chapter is not feasible and thus we provide a short tutorial on how to use a suite of estimators that can easily be accessed via the website “Productivity in R” that can be found as <https://sites.google.com/site/productivityinr/>. We use standard notations in the codes and let n be the number of firms, T the number of time series and nT the total number of observations, i.e., $nT = n \times T$. The data input files configured with the first column a $nT \times 1$ column vector of the dependent variable y , and the next k containing $nT \times 1$ vectors of the independent variables contained in x' . The convention used is for the first T observations to be for the first cross section, the second T observations for the second cross section and so on. The default program output (“results.out” is the default output filename) contains parameter estimates, standard errors, t -values, average technical efficiency, correlation of effects and efficiencies, Spearman rank order correlation of effects and efficiencies, R-squared and adjusted R-squared.

^gWhile originally Simar and Zelenyuk (2011) considered the approach of Kumbhakar et al. (2007) for the first stage, one could use other nonparametric SFA approaches, e.g., a more general version proposed by Park et al. (2015) or Simar et al. (2017).

5.1 Basic model setup

The default model is the panel data model we introduced in [Section 3](#), which we rewrite as follows:

$$y_{it} = \alpha + x'_{it}\beta - u_{it} + v_{it}, \quad (79)$$

where the global mean (α) is subtracted out in the regression and where this demeaning method is applied to the suite of estimators discussed in SS and CSS as well as for the [Battese and Coelli \(1992\)](#) (BC) estimator. Time trends can be added as well to account for disembodied technical change that is available for adoption by all firms and various right-hand-side (rhs) variables that may be correlated with the effects can be identified in the main R file. Efficiencies can also be averaged over different estimators using model averaging weights based on a simple average ($\text{AVE}=0$), AIC weights ($\text{AVE}=1$), or BIC weights ($\text{AVE}=2$). Outliers can be addressed by trimming (this does not apply to the BC and the [DEA](#) estimators).

5.2 Figures and tables

Figures and tables are selected by setting values of 1 to print and 0 to skip the results. There are a number of figures that are already set up to print. For example, when `fig1=1` a figure for the average of efficiencies of the time-variant estimators is printed. When `fig2=1` a figure for the efficiencies from the time-variant estimators is printed. When `fig3=1` a figure for the average of efficiencies from all estimators is printed, while `fig4=1` the weighted average of the efficiencies is printed. `tab1=1` prints a table for the average of efficiencies from the time-variant estimators, `tab2=1` prints one for the efficiencies from the time-variant estimators are printed and `tab3=1` prints a table of the individual effects from all estimators. Also, setting `firmeff=1` saves the efficiencies of each individual firm for each estimator utilized.

5.3 Different estimators

The R codes for the SFA models include the fixed effect, random effects, and the Hausman–Taylor (`HT`) version of the stochastic panel frontier models of [Schmidt and Sickles \(1984\)](#). The `FR` and the `HT` global parameters are set to determine which model is estimated. Since the `HT` estimator of the panel stochastic frontier allows for selected regressors to be correlated with the efficiency effects term the global option `k1` is set at the number of variables in X not correlated with the efficiency effects and are loaded into the first `k1` columns for the regressors. These different models are referred to as `FIX`, `RND`, and `HT`. The global options `PSS1`, `PSS2`, and `PSS3` designate the estimators for the [Park et al. \(1998, 2003, 2007\)](#) models. The value of 1 prints the results

and 0 skips the results. Bandwidth selection is based on leave-one-out least squares cross-validation. The `CSS` global option allows for the estimates of the [Cornwell et al. \(1990\)](#) models to be generated. A global option of `CSS=1` prints results of the fixed effects/within estimator (`CSSW`) estimator, while `CSS=2` generates the GLS random effects estimator. The efficient IV estimator (analogous to the HT estimator for the SS Model) is engaged with `CSS=3`. Finally, `CSS=4` prints out all four model results. Time-invariant variables should be placed at the end of X . The parameter `zp` is set equal to the number of time-invariant variables. The [Kneip et al. \(2012\)](#) factor model is called by setting the global parameter `KSS=1`. The `KSS` estimator also finds the number of factors and the code varies the number of factors from L_{max} to L_{min} and finds the first highest number at which the dimensionality test is not rejected. Cross-validation is used to estimate the optimal smoothing parameter. `gr_st` is the starting point of the grid search, `gr_in` is the increment in the grid search, and `gr_en` is the end point of the grid search. The [Battese and Coelli \(1992\)](#) model is called by setting the `BC` option=1. The bounded inefficiency model of [Almanidis et al. \(2014\)](#) is called by setting the `BIE` option=1. Different distributions for the lower bound on inefficiency can also be specified using the `bie_dist` parameter. For a `bie_dist=0` the truncated exponential is specified, while `bie_dist=1` uses the truncated half-normal, and `bie_dist=2` the doubly truncated normal distribution. Finally, the [Kutlu \(2018\)](#) endogeneity correction for selected regressors (e.g., input levels in the production frontier) based on the CSS estimator is also available on this website along with instructions for its use.

6 DEA efficiency software in R

The [Jeon and Sickles \(2004\)](#) model is the directional distance function method outlined in the section on DEA estimators. This estimator also allows us to address the presence of undesirable outputs. Jeon and Sickles (JS) used this DEA estimator to examine the productivity effects of controlling for carbon dioxide emissions on productivity growth using Malmquist indexes. The JS estimator uses OECD data while Efficiency software uses UNIDO and Bank data. The JS R codes are in a separate `Jeon_Sickles_2004.zip` file on the website. The default results that are printed by the R codes are productivity growth, efficiency change, and technology change and confidence intervals for the growth decompositions are based on the bootstrapping methods discussed in [Simar and Wilson \(2007\)](#) and the references therein. Next, we have the [Simar and Zelenyuk \(2006\)](#) model, which implements the [Li \(1996\)](#) test in the context of comparing distributions of efficiency estimated via DEA and the [Simar and Zelenyuk \(2007\)](#) model that constructs confidence intervals and bias corrections for DEA-estimated aggregate efficiencies of a set of firms. It also provides a test for the comparison of these group efficiencies.

There are other open source R code platforms other than the one that we have focused on in our discussions so far. The *Benchmarking* package in R prepared by Peter Bogetoft and Lars Otto is one such program and the manual and other directions for its use can be found at <https://cran.r-project.org/web/packages/Benchmarking/Benchmarking.pdf>. This can be used to estimate both DEA and FDH models. There have been many other programs developed by Simar and his colleagues in MATLAB® and these can be compiled into R using MATLAB-to-R converters. One such package, the “matconv” package by Siddarta Jairam is a useful conversion tool, although some testing for accuracy is always advised for such automatic translations.^h Many of these MATLAB® programs can be found at <https://sites.google.com/site/productivityefficiency>ⁱ. Finally, Daraio et al. (2018a) provide a survey on a variety of software platforms that can estimate SFA and DEA models, including a number with R coding.

7 Summary and final remarks

This chapter has discussed a number of statistical and programming techniques used to evaluate the production efficiency of economic units, whether they be firms, sectors, or countries. We have discussed the methods most widely used for these sorts of evaluations and have provided references and URLs to the most up to date websites that provide the R code to implement these methods. We trust that the interested reader will find our discussions and the software helpful for the purposes of practical evaluations of the performance of business entities as well as in their academic research. We have also provided an up to date set of references that productivity and efficiency researchers can use for extended and deeper readings on these widely used and adopted performance benchmarking methods.

Acknowledgments

The authors would like to thank Kerda Varaku for her assistance in preparing our manuscript. All remaining errors are the responsibility of the authors. V.Z. acknowledges the financial support from the Australian Research Council grant (ARC FT170100401).

References

- Adams, R.M., Berger, A.N., Sickles, R.C., 1999. Semiparametric approaches to stochastic panel frontiers with applications in the banking industry. *J. Bus. Econ. Stat.* 17 (3), 349–358.
- Afriat, S.N., 1972. Efficiency estimation of production functions. *Int. Econ. Rev.* 13 (3), 568–598.

^hSee <https://cran.r-project.org/web/packages/matconv/matconv.pdf> for more details.

ⁱArne Henningsen has provided the R conversion to a number of Fortran-based programs (FRONTIER) developed by Tim Coelli as well as a set of new programs in his *A Package for SFA in R*. This platform is available at <https://cran.r-project.org/>.

- Agrell, P.J., Bogetoft, P., et al., 2017. Regulatory benchmarking: models, analyses and applications. *Data Envelopment Anal.* 3 (1–2), 49–91.
- Ahn, S.C., Lee, Y.H., Schmidt, P., 2007. Stochastic frontier models with multiple time-varying individual effects. *J. Prod. Anal.* 27 (1), 1–12.
- Ahn, S.C., Lee, Y.H., Schmidt, P., 2013. Panel data models with multiple time-varying individual effects. *J. Economet.* 174 (1), 1–14.
- Aigner, D., Lovell, C., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. *J. Economet.* 6 (1), 21–37. ISSN 0304-4076. [https://doi.org/10.1016/0304-4076\(77\)90052-5](https://doi.org/10.1016/0304-4076(77)90052-5).
- Allais, M., 1943. *Traité D'Économie Pure*, vol. 3. Imprimerie Nationale, Paris, FR.
- Allen, R., Athanassopoulos, A., Dyson, R.G., Thanassoulis, E., 1997. Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions. *Ann. Oper. Res.* 73, 13–34.
- Almanidis, P., Sickles, R.C., 2012. The skewness problem in stochastic frontier models: fact or fiction? In: Van Keilegom, I., Wilson, P.W. (Eds.), *Exploring Research Frontiers in Contemporary Statistics and Econometrics: A Festschrift in Honor of Léopold Simar*. Springer, New York, NY, pp. 201–227.
- Almanidis, P., Qian, J., Sickles, R.C., 2014. Stochastic frontier with bounded inefficiency. In: Sickles, R.C., Horrace, W.C. (Eds.), *Festschrift in Honor of Peter Schmidt: Econometric Methods and Applications*. Springer, New York, NY, pp. 47–82.
- Bada, O., Liebl, D., 2014. The R-package phtt: panel data analysis with heterogeneous time trends. *J. Stat. Softw.* 59 (6), 1–33.
- Bai, J., 2009. Panel data models with interactive fixed effects. *Econometrica* 77 (4), 1229–1279.
- Bai, J., Ng, S., 2002. Determining the number of factors in approximate factor models. *Econometrica* 70 (1), 191–221.
- Banker, R.D., 1993. Maximum likelihood, consistency and data envelopment analysis: a statistical foundation. *Manage. Sci.* 39 (10), 1265–1273.
- Banker, R.D., Charnes, A., Cooper, W.W., 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Manage. Sci.* 30 (9), 1078–1092.
- Battese, G.E., Coelli, T.J., 1992. Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India. *J. Prod. Anal.* 3 (1–2), 153–169.
- Battese, G.E., Coelli, T.J., 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empir. Econ.* 20 (2), 325–332.
- Bauer, P.W., Hancock, D., 1993. The efficiency of the Federal Reserve in providing check processing services. *J. Bank. Financ.* 17 (2–3), 287–311.
- Bloom, N., Genakos, C., Sadun, R., Van Reenen, J., 2012. Management practices across firms and countries. *Acad. Manage. Perspect.* 26 (1), 12–33.
- Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., Roberts, J., 2013. Does management matter? Evidence from India. *Q. J. Econ.* 128 (1), 1–51.
- Bogetoft, P., 1996. DEA on relaxed convexity assumptions. *Manage. Sci.* 42 (3), 457–465.
- Bogetoft, P., 2013. *Performance Benchmarking: Measuring and Managing Performance*. Springer Science & Business Media.
- Chambers, R., Färe, R., Grosskopf, S., 1996. Productivity growth in APEC countries. *Pac. Econ. Rev.* 1 (3), 181–190.
- Charnes, A., Cooper, W., Rhodes, E., 1978. Measuring the efficiency of decision making units. *Eur. J. Oper. Res.* 2 (6), 429–444.
- Charnes, A., Cooper, W.W., Huang, Z.M., Sun, D.B., 1990. Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. *J. Economet.* 46 (1–2), 73–91.

- Chowdhury, H., Zelenyuk, V., 2016. Performance of hospital services in Ontario: DEA with truncated regression approach. *Omega* 63, 111–122.
- Chung, Y.H., Färe, R., Grosskopf, S., 1997. Productivity and undesirable outputs: a directional distance function approach. *J. Environ. Manage.* 51 (3), 229–240.
- Colombi, R., Martini, G., Vittadini, G., 2011. A stochastic frontier model with short-run and long-run inefficiency random effects (Working Paper No. 012011). University of Bergamo, Department of Economics and Technology Management.
- Colombi, R., Kumbhakar, S.C., Martini, G., Vittadini, G., 2014. Closed-skew normality in stochastic frontiers with individual effects and long/short-run efficiency. *J. Prod. Anal.* 42 (2), 123–136.
- Cornwell, C., Schmidt, P., Sickles, R.C., 1990. Production frontiers with cross-sectional and time-series variation in efficiency levels. *J. Econom.* 46 (1), 185–200.
- Curi, C., Lozano-Vivas, A., Zelenyuk, V., 2015. Foreign bank diversification and efficiency prior to and during the financial crisis: does one business model fit all? *J. Bank. Financ.* 61 (S1), S22–S35.
- Dakpo, K.H., Jeanneaux, P., Latruffe, L., 2017. Modelling pollution-generating technologies in performance benchmarking: recent developments, limits and future prospects in the nonparametric framework. *Eur. J. Oper. Res.* 250 (2), 347–359.
- Daraio, C., Kerstens, K.H., Nepomuceno, T.C.C., Sickles, R., 2018a. Productivity and efficiency analysis software: an exploratory bibliographical survey of the options. *J. Econ. Surv.* 1–16 <https://doi.org/10.1111/joes.12270>.
- Daraio, C., Simar, L., Wilson, P.W., 2018b. Central limit theorems for conditional efficiency measures and tests of the “separability” condition in nonparametric, two-stage models of production. *Econom.* 21 (2), 170–191 <https://doi.org/10.1111/ectj.12103>.
- Debreu, G., 1951. The coefficient of resource utilization. *Econometrica* 19 (3), 273–292.
- Demchuk, P., Zelenyuk, V., 2009. Testing differences in efficiency of regions within a country: the case of Ukraine. *J. Prod. Anal.* 32 (2), 81–102.
- Deprins, D., Simar, L., Tulkens, H., 1984. Measuring labour efficiency in post offices. In: Marchand, M., Pestieau, P., Tulkens, H. (Eds.), *The Performance of Public Enterprises: Concepts and Measurement*. Springer, Amsterdam, NL, pp. 243–267.
- DeYoung, R., 1998. Management quality and X-inefficiency in national banks. *J. Financ. Serv. Res.* 13 (1), 5–22.
- Diewert, W.E., 1983. The measurement of waste within the production sector of an open economy. *Scand. J. Econ.* 85 (2), 159–179.
- Du, K., Worthington, A.C., Zelenyuk, V., 2018. Data envelopment analysis, truncated regression and double-bootstrap for panel data with application to Chinese banking. *Eur. J. Oper. Res.* 265 (2), 748–764. ISSN 0377-2217. <https://doi.org/10.1016/j.ejor.2017.08.005>.
- Dyson, R.G., Thanassoulis, E., 1988. Reducing weight flexibility in data envelopment analysis. *J. Oper. Res. Soc.* 39 (6), 563–576.
- Färe, R., Grosskopf, S., 1983. Measuring congestion in production. *Z. National. J. Econ.* 43 (3), 257–271.
- Färe, R., Grosskopf, S., 1996. *Intertemporal Production Frontiers: With Dynamic DEA*. Kluwer Academic Publishers, Norwell, MA.
- Färe, R., Grosskopf, S., 2003. Nonparametric productivity analysis with undesirable outputs: comment. *Am. J. Agric. Econ.* 85 (4), 1070–1074.
- Färe, R., Grosskopf, S., 2004. Modeling undesirable factors in efficiency evaluation: comment. *Eur. J. Oper. Res.* 157 (1), 242–245.

- Färe, R., Grosskopf, S., 2009. A comment on weak disposability in nonparametric production analysis. *Am. J. Agric. Econ.* 91 (2), 535–538.
- Färe, R., Lovell, C.A.K., 1978. Measuring the technical efficiency of production. *J. Econ. Theory* 19 (1), 150–162. ISSN 0022-0531. [https://doi.org/10.1016/0022-0531\(78\)90060-1](https://doi.org/10.1016/0022-0531(78)90060-1).
- Färe, R., Svensson, L., 1980. Congestion of production factors. *Econom. J. Econom. Soc.* 48 (7), 1745–1753.
- Färe, R., Grosskopf, S., Logan, J., 1983. The relative efficiency of Illinois electric utilities. *Resour. Energy* 5 (4), 349–367.
- Färe, R., Grosskopf, S., Roos, P., 1996. On two definitions of productivity. *Econ. Lett.* 53 (3), 269–274.
- Färe, R., Grosskopf, S., Zelenyuk, V., 2007. Finding common ground: efficiency indices. In: Färe, R., Grosskopf, S., Primont, D. (Eds.), *Aggregation, Efficiency and Measurement*. Springer, Boston, MA, pp. 83–95.
- Farrell, M.J., 1957. The measurement of productive efficiency. *J. R. Stat. Soc. Ser. A (General)* 120 (3), 253–290.
- Førsund, F.R., 2009. Good modelling of bad outputs: pollution and multiple-output production. *Int. Rev. Environ. Resour. Econ.* 3 (1), 1–38.
- Foster, L., Haltiwanger, J., Syverson, C., 2008. Reallocation, firm turnover, and efficiency: selection on productivity or profitability? *Am. Econ. Rev.* 98 (1), 394–425.
- Frantz, R.S., 1997. X-Efficiency: Theory, Evidence and Applications, second ed. Kluwer Academic Publishers, Norwell, MA.
- Frantz, R.S., 2007. Empirical evidence on X-efficiency, 1967-2004. In: Frantz, R. (Ed.), *Renaissance in Behavioral Economics: Essays in Honour of Harvey Leibenstein*. Routledge, New York, NY, pp. 221–227.
- Fu, X.M., Heffernan, S., 2009. The effects of reform on China's bank structure and performance. *J. Bank. Financ.* 33 (1), 39–52.
- Gijbels, I., Mammen, E., Park, B.U., Simar, L., 1999. On estimation of monotone and concave frontier functions. *J. Am. Stat. Assoc.* 94 (445), 220–228.
- Glaister, K.W., 2014. The contribution of management to economic growth: a review. *Prometheus* 32 (3), 227–244.
- Greene, W.H., 1980a. Maximum likelihood estimation of econometric frontier functions. *J. Economet.* 13 (1), 27–56.
- Greene, W.H., 1980b. On the estimation of a flexible frontier production model. *J. Econom.* 13 (1), 101–115.
- Greene, W.H., 2005a. Fixed and random effects in stochastic frontier models. *J. Prod. Anal.* 23 (1), 7–32.
- Greene, W.H., 2005b. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *J. Econom.* 126 (2), 269–303.
- Grifell-Tatjé, E., Lovell, C.A.K., 2013. Advances in cost frontier analysis of the firm. In: Thomas, C.R., Shughart, W. (Eds.), *The Oxford Handbook of Managerial Economics*. Oxford University Press, London, UK, pp. 1–18.
- Grifell-Tatjé, E., Lovell, C.A.K., 2015. *Productivity Accounting*. Cambridge University Press, New York, NY.
- Grifell-Tatjé, E., Lovell, C.A.K., Sickles, R.C., 2018. *The Oxford Handbook of Productivity Analysis*. Oxford University Press.
- Griffin, J.E., Steel, M.F.J., 2007. Bayesian stochastic frontier analysis using WinBUGS. *J. Prod. Anal.* 27 (3), 163–176.

- Grosskopf, S., 1986. The role of the reference technology in measuring productive efficiency. *Econ. J.* 96 (382), 499–513.
- Hall, R.E., Jones, C.I., 1999. Why do some countries produce so much more output per worker than others? *Q. J. Econ.* 114 (1), 83–116.
- Hao, G., Wei, Q.L., Yan, H., 2000. A game theoretical model of DEA efficiency. *J. Oper. Res. Soc.* 51 (11), 1319–1329.
- Hsieh, C.-T., Klenow, P.J., 2009. Misallocation and manufacturing TFP in China and India. *Q. J. Econ.* 124 (4), 1403–1448.
- Jeon, B.M., Sickles, R.C., 2004. The role of environmental factors in growth accounting. *J. Appl. Economet.* 19 (5), 567–591.
- Jeong, S.-O., Simar, L., 2006. Linearly interpolated FDH efficiency score for nonconvex frontiers. *J. Multivar. Anal.* 97 (10), 2141–2161.
- Jiang, C., Yao, S., Zhang, Z., 2009. The effects of governance changes on bank efficiency in China: a stochastic distance function approach. *China Econ. Rev.* 20 (4), 717–731.
- Jondrow, J., Lovell, C.A.K., Materov, I.S., Schmidt, P., 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *J. Econom.* 19 (2–3), 233–238.
- Kao, C., 2009a. Efficiency decomposition in network data envelopment analysis: a relational model. *Eur. J. Oper. Res.* 192 (3), 949–962.
- Kao, C., 2009b. Efficiency measurement for parallel production systems. *Eur. J. Oper. Res.* 196 (3), 1107–1112.
- Kao, C., 2014. Network data envelopment analysis: a review. *Eur. J. Oper. Res.* 239 (1), 1–16.
- Kim, M., Schmidt, P., 2008. Valid tests of whether technical inefficiency depends on firm characteristics. *J. Econom.* 144 (2), 409–427.
- Kneip, A., Park, B.U., Simar, L., 1998. A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econom. Theor.* 14 (6), 783–793.
- Kneip, A., Simar, L., Wilson, P.W., 2008. Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econom. Theor.* 24 (6), 1663–1697. ISSN 02664666.
- Kneip, A., Sickles, R.C., Song, W., 2012. A new panel data treatment for heterogeneity in time trends. *Econom. Theor.* 28 (3), 590–628.
- Kneip, A., Simar, L., Wilson, P.W., 2015. When bias kills the variance: central limit theorems for DEA and FDH efficiency scores. *Econom. Theor.* 31 (2), 394–422.
- Kneip, A., Simar, L., Wilson, P.W., 2016. Testing hypotheses in nonparametric models of production. *J. Bus. Econ. Stat.* 34 (3), 435–456.
- Koopmans, T., 1951a. Activity analysis of production and allocation. Wiley, New York, NY.
- Koopmans, T.C., 1951b. Analysis of production as an efficient combination of activities. In: Koopmans, T.C. (Ed.), *Activity Analysis of Production and Allocation*. vol. 13. Wiley, New York, NY, pp. 33–37.
- Korostelev, A., Simar, L., Tsybakov, A.B., 1995a. Efficient estimation of monotone boundaries. *Ann. Stat.* 23 (2), 476–489.
- Korostelev, A., Simar, L., Tsybakov, A.B., 1995b. On estimation of monotone and convex boundaries. *Publ. Inst. Stat. Univ. Paris* 39 (1), 3–18.
- Kumbhakar, S.C., 1990b. Production frontiers, panel data, and time-varying technical inefficiency. *J. Econom.* 46 (1), 201–211.
- Kumbhakar, S., Schmidt, P., 2016. Endogeneity problems in econometrics, *J. Econom.* 190 (2), 209–374 (special issue).
- Kumbhakar, S.C., Park, B.U., Simar, L., Tsionas, E.G., 2007. Nonparametric stochastic frontiers: a local maximum likelihood approach. *J. Econom.* 137 (1), 1–27. ISSN 0304-4076. <https://doi.org/10.1016/j.jeconom.2006.03.006>.

- Kutlu, L., 2018. A distribution-free stochastic frontier model with endogenous regressors. *Econ. Lett.* 163, 152–154.
- Kwan, S.H., 2006. The X-efficiency of commercial banks in Hong Kong. *J. Bank. Financ.* 30 (4), 1127–1147.
- Lee, Y. H., 1991. Panel Data Models With Multiplicative Individual and Time Effects: Applications to Compensation and Frontier Production Functions (Ph.D. thesis). Michigan State University, East Lansing, MI.
- Lee, Y.H., Schmidt, P., 1993. A production frontier model with flexible temporal variation in technical efficiency. In: Fried, H.O., Lovell, C.A.K., Schmidt, S.S. (Eds.), *The Measurement of Productive Efficiency: Techniques and Applications*. Oxford University Press, New York, NY, pp. 237–255.
- Leibenstein, H., 1966. Allocative efficiency vs. “X-efficiency”. *Am. Econ. Rev.* 56 (3), 392–415.
- Leibenstein, H., 1975. Aspects of the X-efficiency theory of the firm. *Bell J. Econ.* 6 (2), 580–606.
- Leibenstein, H., 1987. *Inside the Firm: The Inefficiencies of Hierarchy*. Harvard University Press, Cambridge, MA.
- Li, Q., 1996. Nonparametric testing of closeness between two unknown distribution functions. *Economet. Rev.* 15 (3), 261–274. <https://doi.org/10.1080/07474939608800355>. <http://www.tandfonline.com/doi/pdf/10.1080/07474939608800355>.
- Liang, L., Wu, J., Cook, W.D., Zhu, J., 2008. The DEA game cross-efficiency model and its Nash equilibrium. *Ope. Res.* 56 (5), 1278–1288.
- Lieberman, M.B., Lau, L.J., Williams, M.D., 1990. Firm-level productivity and management influence: a comparison of US and Japanese automobile producers. *Manage. Sci.* 36 (10), 1193–1215.
- Liu, J., Sickles, R.C., Tsionas, E.G., 2017. Bayesian treatments to panel data models with time-varying heterogeneity. *Econometrics* 5 (33), 1–21.
- Lozano, S., 2012. Information sharing in DEA: a cooperative game theory approach. *Eur. J. Oper. Res.* 222 (3), 558–565.
- Luenberger, D.G., 1992. Benefit functions and duality. *J. Math. Econ.* 21 (5), 461–481.
- Materov, I.S., 1981. On full identification of the stochastic production frontier model. *Ekonomika i Matematicheskie Metody* 17, 784–788.
- Meeusen, W., van den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *Int. Econ. Rev.* 18 (2), 435–444.
- Mester, L.J., 1993. Efficiency in the savings and loan industry. *J. Bank. Financ.* 17 (2–3), 267–286.
- Nakabayashi, K., Tone, K., 2006. Egoist’s dilemma: a DEA game. *Omega* 34 (2), 135–148.
- Nallari, R., Bayraktar, N., 2010. Micro Efficiency and Macro Growth (Working Paper No. 5267). The World Bank Policy Research.
- Olson, J.A., Schmidt, P., Waldman, D.M., 1980. A Monte Carlo study of estimators of the stochastic frontier production function. *J. Econom.* 13 (1), 67–82.
- Park, B.U., Sickles, R.C., Simar, L., 1998. Stochastic panel frontiers: a semiparametric approach. *J. Econom.* 84 (2), 273–301.
- Park, B.U., Simar, L., Weiner, C., 2000. The FDH estimator for productivity efficiency scores: asymptotic properties. *Econom. Theor.* 16 (6), 855–877.
- Park, B.U., Sickles, R.C., Simar, L., 2003. Semiparametric-efficient estimation of AR(1) panel data models. *J. Econom.* 117 (2), 279–309.
- Park, B.U., Sickles, R.C., Simar, L., 2007. Semiparametric efficient estimation of dynamic panel data models. *J. Econom.* 136 (1), 281–301.

- Park, B.U., Simar, L., Zelenyuk, V., 2008. Local likelihood estimation of truncated regression and its partial derivatives: theory and application. *J. Econom.* 146 (1), 185–198. ISSN 0304-4076.
- Park, B.U., Jeong, S.-O., Simar, L., 2010. Asymptotic distribution of conical-hull estimators of directional edges. *Annal. Stat.* 38 (3), 1320–1340.
- Park, B.U., Simar, L., Zelenyuk, V., 2015. Categorical data in local maximum likelihood: theory and applications to productivity analysis. *J. Prod. Anal.* 43 (2), 199–214. ISSN 0895-562X. <https://doi.org/10.1007/s11123-014-0394-y>.
- Petersen, N.C., 1990. Data envelopment analysis on a relaxed set of assumptions. *Manage. Sci.* 36 (3), 305–314.
- Pham, M.D., Zelenyuk, V., 2019. Weak disposability in nonparametric production analysis: A new taxonomy of reference technology sets. *Eur. J. Oper. Res.* 274 (1), 186–198. <https://doi.org/10.1016/j.ejor.2018.09.019>.
- Pham, M.D., Zelenyuk, V., 2018. Slack-based directional distance function in the presence of bad outputs: theory and application to Vietnamese banking. *Empir. Econ.* 54 (1), 153–187.
- Pitt, M.M., Lee, L.-F., 1981. The measurement and sources of technical inefficiency in the Indonesian weaving industry. *J. Dev. Econ.* 9 (1), 43–64.
- Podinovski, V.V., 2015. DEA models with production trade-offs and weight restrictions. In: Zhu, J. (Ed.), *Data Envelopment Analysis: A Handbook of Models and Methods*. Springer, New York, NY, pp. 105–144.
- Podinovski, V.V., Bouzidine-Chameeva, T., 2013. Weight restrictions and free production in data envelopment analysis. *Oper. Res.* 61 (2), 426–437.
- Podinovski, V.V., Kuosmanen, T., 2011. Modelling weak disposability in data envelopment analysis under relaxed convexity assumptions. *Eur. J. Oper. Res.* 211 (3), 577–585.
- Qian, J., Sickles, R.C., 2008. Stochastic Frontiers With Bounded Inefficiency (unpublished manuscript). Shanghai Jiao Tong University, Shanghai, CN.
- Rezvanian, R., Ariss, R.T., Mehdian, S.M., 2011. Cost efficiency, technological progress and productivity growth of Chinese banking pre-and post-WTO accession. *Appl. Financ. Econ.* 21 (7), 437–454.
- Schmidt, P., Sickles, R.C., 1984. Production frontiers and panel data. *J. Bus. Econ. Stat.* 2 (4), 367–374.
- Seiford, L.M., Zhu, J., 2002. Modeling undesirable factors in efficiency evaluation. *Eur. J. Oper. Res.* 142 (1), 16–20.
- Shephard, R.W., 1953. *Cost and Production Functions*. Princeton University Press, Princeton, NJ.
- Shephard, R.W., 1970. *Theory of Cost and Production Functions*. Princeton Studies in Mathematical Economics. Princeton University Press. Princeton, NJ.
- Shephard, R.W., 1974. Indirect production functions. In: *Mathematical Systems in Economics*, vol. 10. Anton Hain, Meisenheim am Glan, Germany.
- Sickles, R., Zelenyuk, V., 2019. *Measurement of Productivity and Efficiency: Theory and Practice*. Cambridge University Press, New York, NY (forthcoming).
- Simar, L., 2007. How to improve the performances of DEA/FDH estimators in the presence of noise? *J. Prod. Anal.* 28 (3), 183–201. ISSN 0895-562X. <https://doi.org/10.1007/s11123-007-0057-3>.
- Simar, L., Wilson, P.W., 2007. Estimation and inference in two-stage, semi-parametric models of production processes. *J. Econom.* 136 (1), 31–64.
- Simar, L., Wilson, P.W., 2008. Statistical inference in nonparametric frontier models: recent developments and perspectives. In: Fried, H.O., Knox Lovell, C.A., Schmidt, S.S. (Eds.), *The Measurement of Productive Efficiency*. Oxford University Press, New York, NY, pp. 421–521.

- Simar, L., Wilson, P.W., 2011. Two-stage DEA: caveat emptor. *J. Prod. Anal.* 36 (2), 205–218.
- Simar, L., Wilson, P.W., 2013. Estimation and inference in nonparametric frontier models: recent developments and perspectives. *Found. Trends Econom.* 5 (3-4), 183–337.
- Simar, L., Zelenyuk, V., 2006. On testing equality of distributions of technical efficiency scores. *Economist. Rev.* 25 (4), 497–522. <https://doi.org/10.1080/07474930600972582>. <http://www.tandfonline.com/doi/pdf/10.1080/07474930600972582>.
- Simar, L., Zelenyuk, V., 2007. Statistical inference for aggregates of Farrell-type efficiencies. *J. Appl. Econom.* 22 (7), 1367–1394.
- Simar, L., Zelenyuk, V., 2011. Stochastic FDH/DEA estimators for frontier analysis. *J. Prod. Anal.* 36 (1), 1–20. ISSN 0895-562X. <https://doi.org/10.1007/s11123-010-0170-6>.
- Simar, L., Zelenyuk, V., 2018. Central limit theorems for aggregate efficiency. *Oper. Res.* 166 (1), 139–149.
- Simar, L., Van Keilegom, I., Zelenyuk, V., 2017. Nonparametric least squares methods for stochastic frontier models. *J. Prod. Anal.* 47 (3), 189–204.
- Stevenson, R.E., 1980. Likelihood functions for generalized stochastic frontier estimation. *J. Econom.* 13 (1), 57–66.
- Sueyoshi, T., Yuan, Y., Goto, M., 2017. A literature study for DEA applied to energy and environment. *Energy Econ.* 62, 104–124.
- Thompson, R.G., Langemeier, L.N., Lee, C.-T., Lee, E., Thrall, R.M., 1990. The role of multiplier bounds in efficiency analysis with application to Kansas farming. *J. Econom.* 46 (1–2), 93–108.
- Tsionas, E.G., Kumbhakar, S.C., 2014. Firm heterogeneity, persistent and transient technical inefficiency: a generalized true random-effects model. *J. Appl. Econom.* 29 (1), 110–132.
- Tsionas, E.G., Papadakis, E.N., 2010. A Bayesian approach to statistical inference in stochastic DEA. *Omega* 38 (5), 309–314.
- Tyteca, D., 1996. On the measurement of the environmental performance of firms—a literature review and a productive efficiency perspective. *J. Environ. Manage.* 46 (3), 281–308.
- von Neumann, J., 1945. A model of general equilibrium. *Rev. Econ. Stud.* 13 (1), 1–9.
- Wang, H.-J., Schmidt, P., 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. *J. Prod. Anal.* 18 (2), 129–144.
- Yao, S., Han, Z., Feng, G., 2008. Ownership reform, foreign competition and efficiency of Chinese commercial banks: a non-parametric approach. *World Econ.* 31 (10), 1310–1326.
- Zelenyuk, V., Zheka, V., 2006. Corporate governance and firm's efficiency: the case of a transitional country, Ukraine. *J. Prod. Anal.* 25 (1), 143–157.

Chapter 9

Stochastic frontier models using R[☆]

Giancarlo Ferrara*

Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy

SOSE—Soluzioni per il Sistema Economico SpA, Rome, Italy

*Corresponding author: e-mail: giancarlo.ferrara@gmail.com

Abstract

The production function is usually assumed to specify the maximum output obtainable, from a given set of inputs, describing the boundary or frontier of the obtainable output from each feasible combination of input; it relates the production process of individual units to the efficient border of the production possibilities. The measure of the distance of each unit from the border is the most immediate way to assess its (in)efficiency. However, the production function is not generally known, but it has only a set of information on each production unit and it is therefore essential to develop techniques to estimate the production frontier. Starting from the packages already developed in the R environment, this work introduces the methodological aspects of the stochastic frontier models, including a brief introduction to the relative extensions in presence of contextual variables and spatial external factors, comparing the standard stochastic frontier analysis and the semiparametric one. Some simulation studies and an empirical application to agricultural data illustrate the different techniques.

Keywords: Stochastic frontier, Semiparametric, Generalized additive model, Splines, Efficiency, R, Agriculture

1 Introduction

The distance between the output obtained and the corresponding maximum amount obtainable is the most immediate way to measure (in)efficiency (Farrell, 1957). The study of the determinants of efficiency is therefore based on the knowledge of the so-called production function because it makes possible the relation of the production process of individual units to the *efficient*

[☆]The views expressed in the article are those of the author and do not involve the responsibility of SOSE SpA.

border of the production possibilities; indeed, the production function defines the maximum output obtainable from a set of inputs with a given technology.

However, the production function is not generally known, but it has only a set of information on each production unit, and it is essential to develop techniques to estimate the frontier itself.

In a parametric framework it is essential to specify a priori an explicit functional form of the boundary of the production set. The parametric model with stochastic production frontier, as introduced by [Aigner et al. \(1977\)](#) and [Meeusen and van den Broeck \(1977\)](#), lets to analyze the sources of (in)efficiency that are also not directly attributable to the production function or to accidental noise (not directly attributable to the single unit) thus defining a model with a compound error term, noise + (in)efficiency.

Despite its limited computational complexity, the stochastic frontier approach has an important drawback: the lack of flexibility associated with the specification of the production function. Indeed, the assumptions about the functional form of the frontier are often too restrictive and not always appropriated: this issue can introduce substantial bias and might lead to misleading conclusions about, in general terms, the entire production analysis.

To overcome some of the limitations associated with the stochastic frontier production model in its fully parametrized formulation, several semi- and nonparametric approaches have been introduced in literature. In a cross-sectional setting, [Fan et al. \(1996\)](#) introduce a two-step pseudo-likelihood procedure for the estimation of stochastic frontier model where the functional form of the frontier is not known and estimated via kernel regression; more recently, [Kumbhakar et al. \(2007\)](#) propose a novel approach based on the local maximum likelihood principle.

Starting from the link between the study of the stochastic frontier and the relative conditional expectation of the response (output) variable, [Vidoli and Ferrara \(2015\)](#) extend, in a cross-sectional setting, the [Fan et al. \(1996\)](#) approach by considering the specification of the production frontier by means of generalized additive models (GAM—[Hastie and Tibshirani, 1990](#)). [Ferrara and Vidoli \(2017\)](#) further generalize their previous work introducing a semi-parametric model that yields the opportunity of introducing the effect of potential contextual variables, that is the inclusion of exogenous factors that may exert an influence on the producer performance by directly affecting the production process itself or directly the efficiency with which the production process is operating. Moreover, they introduce a penalty on the estimation procedure (if required) that lets to respect the monotonicity constraint between each input and the corresponding output; this property is achieved by using the B-splines approach with penalties, known as P-splines ([Eilers and Marx, 1996](#)), for the nonparametric modeling of the relevant GAM.

The rest of this work is organized as follows. Starting from the packages already developed in the R environment, [Section 2](#) is devoted to the methodological aspects of the stochastic frontier models, including a brief introduction to the relative extensions in presence of contextual variables and spatial external factors. With the aim of comparing the standard stochastic frontier analysis

and the semiparametric one, [Section 3](#) provides some simulation studies, whereas [Section 4](#) provides an empirical application to agricultural data. [Section 5](#) concludes this chapter.

2 Methods

In econometric literature, specification and estimation of production frontier functions are usually carried out by two different approaches: stochastic frontier analysis (SFA) and data envelopment analysis (DEA), respectively. In an SFA analysis it is essential to specify a priori an explicit functional form of the boundary of the production set, while the DEA approach is characterized by the ability to determine the relative efficiency of such units through linear programming, without specifying any functional form for the production function. In other words, unlike parametric techniques, the DEA allows for the determination of the relative efficiency in the absence of a similar detailed description of the production process. If the latter seems to make this approach particularly flexible and generalizable, the main drawback of DEA is its deterministic nature. When using this procedure it is not possible to recognize if the difference in efficiency, namely the distance between observed and maximum possible output, are due to technical inefficiency or effects of disturbance of an accidental type ([Greene, 2008](#)). Therefore, it is not possible to determine whether, for example, inefficiency is due to an adverse condition of the contextual factors, therefore independent of the actions of the entrepreneur, for instance, a season of no rain will not help farm managers get good results, or it can be expressed as the determinant of other factors, such as the quality of personnel management within the enterprise.

The parametric model with stochastic production frontier in addition to providing useful information on the productive asset, exceeds the limits associated with the above model, resulting in an entire analysis of the sources of inefficiency that are not directly attributable to the production function or disturbances of an accidental type, and therefore are not directly attributable to corporate *policy*. The most important drawback associated with the SFA approach is the lack of flexibility associated with the specification of the production function.

In a cross-sectional setting, the conventional stochastic frontier model of [Aigner et al. \(1977\)](#) and [Meeusen and van den Broeck \(1977\)](#) can be written as

$$y_i = f(\mathbf{x}_i; \boldsymbol{\beta}) + v_i - u_i, \quad i = 1, \dots, n, \quad (1)$$

where $Y_i \in \mathbb{R}_+$ is the single output of unit i , $\mathbf{X}_i \in \mathbb{R}_+^p$ is the vector of inputs, $f(\cdot)$ defines a production (frontier) relationship between the single output Y and inputs X depending by the corresponding parameter vector $\boldsymbol{\beta}$, v_i is a symmetric two-sided error representing random effects and $u_i > 0$ is one-sided error term which represents technical inefficiency; $f(\mathbf{x}_i; \boldsymbol{\beta}) + v_i$ is the stochastic frontier. Following common practice, v and u are each identically

independently distributed (iid), $v \sim N(0, \sigma_v^2)$ and u distributed half-normally on the nonnegative part of the real number line: $u \sim N^+(0, \sigma_u^2)$.

However, forcing $f(\cdot)$ to belong to a parametric family of functions (Translog, Cobb–Douglas) can be too restrictive, even inappropriate, and this may lead to a serious modeling bias and therefore misleading conclusions ([Giannakas et al., 2003](#)).

To overcome drawbacks due to the specification of a particular production function, [Ferrara and Vidoli \(2017\)](#) and [Vidoli and Ferrara \(2015\)](#) propose a GAM framework for the estimation of stochastic production frontier models.

A GAM fits a response variable Y using a sum of smooth functions of the explanatory variables, X_j for $j = 1, \dots, p$. In a regression context with Normal response, the model is

$$\mu = E(Y|X = \mathbf{x}) = \alpha + \sum_{j=1}^p s_j(X_j), \quad (2)$$

where the $s_j(\cdot)$'s are smooth functions ([Hastie and Tibshirani, 1990](#)) standardized so that $E[s_j(X_j)] = 0$. GAMs can provide useful approximations to the regression surface relaxing the linear (polynomial) structure of the additive effects. This additional flexibility alleviates the need to impose a perfect linear relationship between each explanatory variable and the response, yet it explains the variability of the response using an additive function of the inputs as in the corresponding SFA model. Generalized additive models are more flexible and the advantage is that the best transformations are determined simultaneously. One useful feature of additive models is that nonparametric estimators of the unknown functions s_j have one-dimensional convergence rates ([Stone, 1986](#)), which makes them much more accurate than estimating the p -dimensional function, and are able to avoid the “curse of dimensionality.”

Model (1) becomes

$$y_i = \psi(\mathbf{x}_i) + v_i - u_i, \quad i = 1, \dots, n, \quad (3)$$

where the unknown function $\psi(\cdot)$ is modeled via GAMs (2) to relax the linear assumption between inputs and output (represented on log scale), but ensuring the additivity of the input factors; hereafter, we refer to this model as GAM-SFA.

For the estimation of the stochastic frontier model (3) we consider the following two-step procedure as proposed by [Fan et al. \(1996\)](#):

- estimating the conditional expectation $E(Y|X = \mathbf{x})$ (i.e., the “mean” frontier),
- estimating error term parameters (σ_v, σ_u) by [Fan et al. \(1996\)](#) method,

where in the first step a GAM specification is considered for the estimation of $E(Y|X = \mathbf{x})$. This kind of models maintains, on the scale given by the smooth terms s_j 's, the same hypothesis of the corresponding classical SFA model

(Aigner et al., 1977) in terms of additivity of the inputs, separability assumptions and independence between u and v conditionally on X .

After obtaining the “mean” frontier $E(Y|X=\bar{x})$, the estimation of the production function $\psi(\cdot)$ will be achieved by shifting the estimation of the conditional expectation in an amount equal to the average estimate of the expected value of the term of inefficiency (Fan et al., 1996).

Given the parameter estimates, the technical efficiency of each unit is estimated by Jondrow et al. (1982) solution that involves deriving the conditional distribution of the component u respect to the compound error $\varepsilon = v - u$.

The conditional distribution of the component u respect to the compound error $\varepsilon = v - u$ can be written as:

$$f_{u|\varepsilon}(u|\varepsilon) = \frac{1}{\sqrt{2\pi}\sigma^*} \cdot \frac{\exp\left\{-\frac{(u-\varphi^*)^2}{2\sigma^*}\right\}}{\left[1 - \Phi\left(-\frac{\varphi^*}{\sigma^*}\right)\right]},$$

where $\varphi^* = -\sigma_u^2\varepsilon/\sigma^2$ e $\sigma_*^2 = \sigma_u^2\sigma_v^2/\sigma^2$; consequently, denoting $\lambda = \sigma_u/\sigma_v$ and $\sigma^2 = \sigma_u^2 + \sigma_v^2$, a point estimator of u_i is

$$E(u|\varepsilon) = \frac{\sigma\lambda}{1+\lambda^2} \left[\frac{\phi(z)}{1-\Phi(z)} - z \right], z = \frac{\varepsilon\lambda}{\sigma}, \quad (4)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions, respectively; if the response is measured in logs, as usual in frontier models, the relative estimates of the technical efficiency for each unit is obtained by

$$TE_i = \exp\{-\hat{u}_i\}. \quad (5)$$

The λ parameter is an indicator of the relative variability of the two sources of error and thus the relative contribution of v and u on ε . If $\lambda \rightarrow 0$, the model excludes the presence of technical inefficiency, on the other hand, if $\lambda \rightarrow +\infty$, the stochastic frontier model degenerates into a Deterministic Frontier Analysis type (Aigner and Chu, 1968), where every departure from the frontier is due only to technical inefficiency; values greater than 1 highlights that the inefficiency component has greater influence on to the accidental symmetrical component.

Model (3) includes the linear model as a special case, where $s_j(x_j) = \beta_j x_j$ (Aigner et al., 1977) but it is clearly more general, because the s_j 's can be very arbitrary nonlinear functions. In addition, it is still possible to further customize the specification of the production frontier considering a generalization of Eq. (2) in a such way

$$f(x_1, x_2, \dots, x_p) = \alpha + \sum_{i=1}^p s_j(x_j) + \sum_{i=1}^p \sum_{k < p} s_{kj}(x_k, x_j) + \beta z + \dots,$$

by introducing effects due to interactions among covariates or linear terms.

The gradients of the nonparametric model can be interpreted as partial output elasticities and their sum as elasticity of scale (HenningSEN and Kumbhakar, 2009).

An important stream of recent literature of frontier estimation employs convex regression (StoNED, Kuosmanen, 2008, Kuosmanen and Johnson, 2010, and Kuosmanen and Kortelainen, 2012). This approach is based on piecewise linear functions to approximate the true (but unknown) regression function that are the optimal representation of a monotonic increasing and concave function that minimizes the L_2 -norm of the residuals (Kuosmanen, 2008). Even if StoNED framework provide appealing properties and a good economic interpretation of the coefficients (as vector of marginal products of inputs for each unit), the method is still computationally intensive as highlighted by Lee et al. (2013).

In the `semsfa` package,^a the unknown $\psi(\cdot)$ can be modeled using four possible different smoothers: thin plate regression splines with penalty (Wood, 2003, 2017), P-splines (Eilers and Marx, 1996), kernel (Hayfield and Racine, 2008), and loess (Cleveland et al., 1993).

Other nonparametric smoothing techniques can be used to estimate the s_j 's function, including local linear methods or smoothing splines. While such approaches, as penalized regression splines with penalty, loessm, and kernel, in most times yields reliable results, the computational aspect is nontrivial and can exhibit poor performance with moderate/large data sets. In addition, the selected smoother must allow to impose the aforementioned constraint of motonocity between each input and the corresponding output. For these reasons, we suggest the P-spline approach introduced by Eilers and Marx (1996) for the nonparametric modeling of the relevant GAM satisfying, as illustrated in Bollaerts et al. (2006) and Muggeo and Ferrara (2008), the monotonicity constraint by introducing a penalty on the estimation procedure. The regression or basis splines can be seen as a compromise between linear regression and nonparametric regression models.

As illustrated by Currie and Durban (2002), P-splines offer significant computational advantages, since they involve inversion of matrices of order less than the number of data points and so they can allow the analysis of larger data sets. In addition, they (i) conserve moments like the mean and variances of the data, (ii) fit polynomial data exactly, and (iii) show no boundary effects.

This section ends describing the extensions of stochastic frontier models in presence of contextual variables and spatial external factors, while a detailed description of the computational aspects relative to P-spline is reported in the last section.

^aThe `semsfa` package developed in the R Environment exploits the functionality of different packages: `mgcv` (Wood, 2018), `gamlss` (Rigby and Stasinopoulos, 2005), `np` (Hayfield and Racine, 2008), and `stats` (R Development Core Team, 2018).

2.1 Contextual variables

More recently, attention has been paid to the role of exogenous factors Z 's that may influence either the productivity function $f(\cdot)$ or the (in)efficiency term u . There is not a general rule that defines how the Z -variables influence the productivity or (in)efficiency (or both), as shown by the several approaches already proposed in the literature. For instance, Kumbhakar et al. (1991) assume inefficiency effects to be a specific parametric function of a set of explanatory variables the parameters of which are estimated simultaneously applying the maximum likelihood method, Johnson and Kuosmanen (2011) introduce StoNEZD model estimating an average function that includes exogenous factors (again) linearly showing that standard statistical inference from linear regression analysis (e.g., t -tests) can be applied for asymptotic inferences regarding coefficients; see Kumbhakar and Lovell (2000) for a review and Florens et al. (2014) and Mastromarco and Simar (2015) for recent developments. Moreover, as pointed out by Simar and Wilson (2008), the two stage procedures for the analysis of the effects associated to exogenous factors are restricted to situation where the external factors can only affect the probability of efficiency and not the shape of the production process.

To estimate the determinants of technical (in)efficiency, Eq. (1) is often modified to allow the inefficiency term u_i to be linearly dependent on exogenous (or contextual) variables z_i as:

$$u_i = z_i \boldsymbol{\delta} + w_i, \quad (6)$$

where the random variable w is defined by the truncation of the $N(0, \sigma_u^2)$ distribution such that $w_i > -z_i \boldsymbol{\delta}$ and u is a nonnegative truncation of the $N(z_i \boldsymbol{\delta}, \sigma_u^2)$ distribution (Battese and Coelli, 1995). Additionally, $\boldsymbol{\delta}$ is a vector of parameters for the determinants of technical inefficiency.

This specification takes into account the heterogeneity of each individual unit by modeling the mean of the inefficiency term as a function of the contextual z variables. As such, it introduces variables into the model that directly affect production efficiency and hence indirectly influence each unit's total output y_i (Lensink and Meesters, 2014). As also indicated by Latruffe (2010), these determinants have to be considered in order to generate differenting levels of performance even if, typically, such factors, such as regulatory constraints, network characteristics, ownership form, or climatic conditions, are not under the control of the firm's manager.

While it is possible to specify different functional forms for $f(\cdot)$ (e.g., translog), the parametric stochastic frontier model of Battese and Coelli (1995) is, again, often criticized for its lack of flexibility in defining the production technology. Ferrara and Vidoli (2017) introduce a new approach to investigate and better understand the role of exogenous factors in the analysis of the production frontier $f(\cdot)$ by considering the generalized additive models for location, scale, and shape (GAMLSS) of Rigby and Stasinopoulos (2005).

GAMLSS has been introduced with the aim to overcome some limitations associated with the generalized linear models (McCullagh and Nelder, 1989) and generalized additive models (Hastie and Tibshirani, 1990) especially for the specification of the shape of the conditional distribution. More specifically, a GAMLSS (Rigby and Stasinopoulos, 2005) assumes that the response variable $y \sim D(y; \mu, \omega, \tau, v)$ where $D \in \mathcal{D}$ can be any distribution (including highly skew and kurtotic, continuous, and discrete distributions) or which exhibit heterogeneity (e.g., the scale or shape of the distribution of the response variable changes with explanatory variables). GAMLSS is especially indicated to model a response variable affected by heterogeneity, where the scale or the shape (or both) of the distribution of the response variable is function of suitable explanatory variables.

The first two parameters μ and ω usually stand for the location and scale parameters and the remaining parameters represent the shape (skewness and kurtosis) parameters. Each parameter (μ, ω, τ, v) of the distribution is modeled using specific explanatory variables and can be expressed as linear/non-linear parametric functions and/or smoothing functions of the explanatory variables (e.g., cubic splines, penalized splines, lowess) and/or random effects.

Clearly, a GAMLSS reduces to a conventional GAM when the model includes the location parameter as the only distribution parameter to be regressed on the covariates.

To obtain GAMLSS estimates, Rigby and Stasinopoulos (2005) introduced a penalized likelihood approach based on modified versions of the back-fitting algorithm for conventional GAM estimation following this basic principle: in each iteration, back-fitting steps are successively applied to the four distribution parameters, with the submodel fits of previous iterations used as offset values for those parameters not involved in the current back-fitting step. For an exhaustive description of the general GAMLSS model and the relative algorithms of the `gamlss` package, see Stasinopoulos and Rigby (2007) and Stasinopoulos et al. (2006) for a more detailed description of GAMLSS in terms of variable selection and interpretability of the results. For a more general discussion about variable selection and interpretability in the semiparametric regression modeling, please see Racine et al. (2014).

This approach assumes that contextual variables exert an influence on the producer performance by directly affecting the production process itself and not directly the efficiency with which the production process is operating as for Battese and Coelli (1995).

For the estimation of the stochastic frontier model (3) in presence of contextual variables Z the following two-step procedure,^b as proposed for the GAM-SFA specification, is considered:

^bThis feature is not still available in the `semsfa` package, while the Battese and Coelli (1995) estimation procedure is available in the `frontier` package.

- estimating the conditional expectation $E(Y | X = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ (i.e., the “mean” frontier) via GAMLSS,
- estimating error term parameters (σ_v, σ_u) by pseudo-likelihood estimators (Fan et al., 1996),

where in the first step the \mathbf{Z} -variables are introduced in a nonlinear specification of the relative equations system; technical efficiency can be calculated by Jondrow et al. (1982) formula as for the GAM-SFA specification.

More generally, nonparametric methods are useful but the practitioner should be aware of the relative assumptions. GAMLSS framework provides a tool to fit, compare, and check models, so many alternatives have to be fitted and explored before a final model is selected.

2.2 Spatial external factors

In general terms, disregarding spatial aspects of the data may produce inefficient or biased estimates and, consequently, misleading inference (Anselin, 2001).

In presence of spatial data, the error independence assumption of the standard SFA model is typically rejected and the inefficiency error component can no longer be assumed homoscedastic introducing substantial bias and leading to misleading conclusions. For this reason, attention has therefore shifted to models that allow the controlling of heterogeneity due to spatial effects.

To address cross-country, or more generally, spatial territorial differences, many studies have been carried out, adding at each step a single specific analysis dimension without, however, taking into account the impact as a whole and without investigating the spatial and environmental interactions.

Efficiency literature usually considers spatial heterogeneity, which refers to the fact that efficiency levels may differ depending on the location, whereas spatial dependence refers to the correlation between the efficiency level at the farm and the efficiency levels of the neighboring farms. For instance, spatial dependence in technical efficiency can be found because farmers in an area may emulate each other, it may be due to the level of infrastructure in the area or because of the climatic and topographic conditions of the area where the farm is located. All these are unobservable latent variables that may be spatially correlated (Areal et al., 2012).

Spatial heterogeneity in technical efficiency literature is controlled by introducing variables that account for political divisions of the land such as regions, counties, and provinces. Many authors have also proposed taking into account the effects of some contextual variables on efficiency with the aim of limiting the influence of multiple and heterogeneous production models. In such cases, the heterogeneity has been taken into account by simply adding dummy variables related to the territory with the purpose of correcting the “average levels” of a baseline model.

This approach had the advantage of testing directly the relative impact on estimates, but on the other hand did not consider the global spatial effect,

neglecting all the other effects not included or simply difficult to estimate. Previous models, although in different formulations, start from the logical assumption of identifying the unique variables that influence the efficiency from a spatial point of view and including them as contextual variables as illustrated in the previous section.

[Areal et al. \(2012\)](#) consider spatial external factors (natural or artificial) following a different approach: they do not identify *ex ante* a set of determinants, often statistically and economically difficult to detect, but they suggest to split the inefficiency term into two different components: the first one linked to the spatial lag and the second to the decision making units (DMUs) specificities. They incorporate spatial dependence into technical efficiency analysis by using an autoregressive specification in the inefficiency term with a Bayesian procedure involving the use of a Gibbs sampler and two Metropolis–Hastings steps to estimate the spatial dependence of farm efficiency.

The main idea is that spatial dependence refers to how much the level of technical inefficiency of firm i depends on the level of efficiency given by other firms $j (= 1, \dots, n)$, that is (in)efficiency of firms located at i is related to the neighbor DMU j 's performances ($j \neq i$).

[Fusco and Vidoli \(2015\)](#) in the `ssfa` package consider a maximum likelihood estimation procedure associated to the same schema developed by [Areal et al. \(2012\)](#) assuming, for cross-sectional data, the following production frontier model:

$$y_i = f(\mathbf{x}_i; \boldsymbol{\beta}) + v_i - u_i, \quad i = 1, \dots, n, \quad (7)$$

$$y_i = f(\mathbf{x}_i; \boldsymbol{\beta}) + v_i - \left(1 - \rho \sum_i w_{i.}\right)^{-1} \tilde{u}_i, \quad (8)$$

where

$$u_i \sim N^+(0, (1 - \rho \sum_i w_{i.})^{-2} \sigma_u^2),$$

u_i and v_i are independently distributed of each other and of the regressors,

$$\tilde{u}_i \sim N(0, \sigma_{\tilde{u}}^2),$$

$w_{i.}$ is a standardized row of the spatial weights matrix,

ρ is the spatial lag parameter with $\rho \in (0, 1)$.

For other spatial specifications of stochastic frontier models in R see the `spfrontier` package ([Pavlyuk, 2016](#)). For a complete review on how spatial dependence has been recently introduced into stochastic frontier models, see: [Druska and Horrace \(2004\)](#), [Schmidt et al. \(2009\)](#), [Areal et al. \(2012\)](#), [Glass et al. \(2014, 2016\)](#), [Adetutu et al. \(2015\)](#), [Tsionas and Michaelides \(2016\)](#), and [Billé et al. \(2018\)](#).

2.3 P-splines: Computational aspects

P-splines, as introduced by [Eilers and Marx \(1996\)](#), are a parametric approach defined by bases of spline with penalties. They have many interesting

properties, such as flexibility, ease of computation and a connection to smoothing splines, but in the context of the stochastic frontier models their main feature is making possible to impose additional monotonicity and concavity constraints on the fitted function.

The remainder outlines the P-spline framework as described by [Eilers and Marx \(1996\)](#) and highlights how it is possible to impose additional constraints to the fitted function in order to respect monotonicity and concavity properties separately for each covariate.

Let (x_i, Y_i) be the i -th observation ($i = 1, \dots, n$) associated to the response variable Y and to the covariate X . The interest here is aimed to the estimate of a smooth function $\psi(\cdot)$, such that $E[Y_i|x_i] = \mu_i = \psi(x_i)$. The key concept of this approach, fully parameterized, it is to express the unknown relationship $\psi(\cdot)$ through a linear combination of appropriate basis functions and relevant coefficients $\boldsymbol{\delta} = (\delta_1, \dots, \delta_k)^T$, as follows:

$$\mu_i = \psi(x_i) = \delta_1 B_1(x_i) + \delta_2 B_2(x_i) + \dots + \delta_k B_k(x_i) = \mathbf{B}(x_i)^T \boldsymbol{\delta}, \quad (9)$$

where, $\mathbf{B}(x_i)^T$ is the i -th row of the basis matrix \mathbf{B} of dimensions $n \times k$, such as $\boldsymbol{\mu} = \mathbf{B}\boldsymbol{\delta}$. Among the different basis functions that can be used for the construction of the matrix \mathbf{B} , it will be considered the B-spline ([de Boor, 1978](#)). In general terms, a B-spline consists of polynomial pieces connected at specific points called nodes and it depends on the number and corresponding position of the nodes and the degree q of the polynomial; the dimension of the basis k (i.e., the number of columns of the matrix \mathbf{B}) is given by the number of intervals (i.e., number of knots + 1) plus the degree q .

Once defined the matrix \mathbf{B} , the estimation of $\psi(\cdot)$ becomes a standard regression problem with the k (columns) covariates given by B_1, \dots, B_k ; the relevant objective function ([Eilers and Marx, 1996](#)) to minimize is

$$S = \sum_{i=1}^n \left\{ y_i - \sum_{j=1}^k \delta_j B_j(x_i) \right\}^2,$$

or in compact form:

$$\| \mathbf{y} - \mathbf{B}\boldsymbol{\delta} \|,$$

where $\|\cdot\|$ is the Euclidean norm; the solution is $\hat{\boldsymbol{\delta}} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{y}$ and $\hat{\boldsymbol{\mu}} = \mathbf{B}\hat{\boldsymbol{\delta}}$.

If the degree of the polynomial is a minor issue (usually is fixed equal to 3—cubic), the choice of the number and location of the nodes is a matter of greater importance: too many (few) nodes in fact lead to an overfitting (underfitting) of the fitted $\hat{\psi}(\cdot)$; at the same time, the estimate of the function will be affected by their position. To overcome this problem and thus limits the influence resulting from the choice of a given number of nodes and their position, [Eilers and Marx \(1996\)](#) suggest to consider a large number of equally

spaced nodes and to impose constraints on the coefficients to control the possible overfitting; for this reason, they call this approach penalized spline (P-spline). Since the δ coefficients provide a description of the “heights” of the B-spline, [Eilers and Marx \(1996\)](#) introduce a penalty based on the finite difference of coefficients of adjacent B-spline, in the form: $\alpha\{(\delta_2 - \delta_1)^2 + (\delta_3 - \delta_2)^2 + \dots\}$ or $\alpha\{(\delta_1 - 2\delta_2 + \delta_3)^2 + (\delta_2 - 2\delta_3 + \delta_4)^2 + \dots\}$ for differences of first and second order, respectively. In matrix notation, the difference of order d $\Delta^d \boldsymbol{\delta}$ can be written as: $\Delta^d \boldsymbol{\delta} = \mathbf{D}_d \boldsymbol{\delta}$ where \mathbf{D}_d is the known matrix of the d -difference operator. The penalty then becomes $\alpha \boldsymbol{\delta}^T \mathbf{D}_d^T \mathbf{D}_d \boldsymbol{\delta}$ and the function to be minimized:

$$\|\mathbf{y} - \mathbf{B}\boldsymbol{\delta}\|^2 + \alpha \|\mathbf{D}_d \boldsymbol{\delta}\|^2,$$

with explicit solution (function of α) given by:

$$\hat{\boldsymbol{\delta}}_\alpha = (\mathbf{B}^T \mathbf{B} + \alpha \mathbf{D}_d^T \mathbf{D}_d)^{-1} \mathbf{B}^T \mathbf{y};$$

and the covariance matrix of the estimates is given instead by:

$$(\mathbf{B}^T \mathbf{B} + \alpha \mathbf{D}_d^T \mathbf{D}_d)^{-1} \mathbf{B}^T \mathbf{B} (\mathbf{B}^T \mathbf{B} + \alpha \mathbf{D}_d^T \mathbf{D}_d)^{-1}.$$

Remembering that $\alpha > 0$ influence the smoothness of the fitted curve, it can be noted that for low values of α $\hat{\psi}(\cdot)$ will be more irregular, for high values of α the estimated curve will be more regular until $\alpha \rightarrow +\infty$ where the curve will approximate a polynomial of degree $d - 1$, with d the order of penalty used.

Within this framework it is necessary to define additional criteria to choose the better value for α . Among the different popular indexes (e.g., Akaike information criterion (AIC), BIC, Mallow's C_p and cross-validation) the most used is the generalized cross-validation given by: $GCV(\alpha) = \|\mathbf{y} - \hat{\boldsymbol{\mu}}_\alpha\|^2 \{1 - n^{-1} \text{tr}(\mathbf{S}_\alpha)\}^{-2}$, where tr is the trace operator and \mathbf{S}_α is the hat-matrix of the model, such as: $\hat{\boldsymbol{\mu}}_\alpha = \mathbf{S}_\alpha \mathbf{y}$; note that the trace of the matrix \mathbf{S}_α gives a measure of the effective model dimension that will reduce to k when $\alpha \rightarrow +\infty$.

In the B- or P-spline framework, it is possible to impose monotonicity constraints on the estimated function $\psi(\cdot)$, as described in [Bollaerts et al. \(2006\)](#). Indeed, [de Boor \(1978\)](#) provides the solution for the first derivative of the function $\psi(\cdot)$ specified by B-spline basis with equally spaced nodes:

$$\psi'(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} \sum_{k=1}^K \delta_k \mathbf{B}_k(\mathbf{x}; q) = h^{-1} \sum_{k=1}^{K-1} \Delta^1 \delta_k \mathbf{B}_k(\mathbf{x}; q-1), \quad (10)$$

where h is the distance among nodes and $\{\mathbf{B}_k(\mathbf{x}, q-1)\}_{k=1}^{K-1}$ the functions relative to a B-spline basis of order $q - 1$. Since h and the relative basis are positive, it will be sufficient to impose the constraint that $\Delta^1 \delta_j$ has to be positive (negative) to impose the relative positivity (negativity) of $\psi'(\cdot)$. With regard to the stochastic frontier, the idea is to penalize those values of $\Delta^1 \delta_j$ less desirable, that is violation of the condition $\Delta^1 \delta_j \geq 0$.

This result may be achieved by including a further asymmetric penalty about the first-order differences, given by: $\sum_{k=1}^{K-1} I(\Delta^1 \delta_k < 0) \Delta^1 \delta_k$ (or $I(\Delta^1 \delta_k \geq 0)$), where $I(\cdot)$ is the indicator function of event. If the k -th difference ($\Delta^1 \delta_k$) is greater or equal to zero (less or equal), monotonicity is ensured and no penalty is required; otherwise, if $\Delta^1 \delta_k < 0$ (or $\Delta^1 \delta_k > 0$) it is necessary strongly penalize such violation weighting it through a very high quantity, for instance $\kappa_1 = 10^6$ (Muggeo and Ferrara, 2008). In matrix notation, the penalty is $\kappa_1 \boldsymbol{\delta}^T \mathbf{D}_1^T \mathbf{V}_1 \mathbf{D}_1 \boldsymbol{\delta}$ with $\mathbf{V}_1 = \text{diag}(I(\Delta^1 \delta_1 < 0), \dots, I(\Delta^1 \delta_{K-1} < 0))$, the new objective function to minimize:

$$\| \mathbf{y} - \mathbf{B}\boldsymbol{\delta} \|^2 + \alpha \| \mathbf{D}_d \boldsymbol{\delta} \|^2 + \kappa_1 \| \mathbf{V}_1^{1/2} \mathbf{D}_1 \boldsymbol{\delta} \|^2,$$

and the parameter estimates given by:

$$\hat{\boldsymbol{\delta}}_\alpha = (\mathbf{B}^T \mathbf{W} \mathbf{B} + \alpha \mathbf{D}_d^T \mathbf{D}_d + \kappa_1 \mathbf{D}_1^T \mathbf{V}_1 \mathbf{D}_1)^{-1} \mathbf{B}^T \mathbf{W} \mathbf{y}.$$

As earlier mentioned, forcing the finite first-order difference of adjacent coefficients to be positive is a sufficient condition for the first-order derivative of the B-spline function to be strictly positive. Similarly, by induction, Bollaerts et al. (2006) show that a constraint of positivity (negativity) on second-order finite differences of the coefficients, constitutes a sufficient condition for the second derivative B-spline function to be positive (negative). Therefore, to impose a constraint of concavity (convexity) on the function $\psi(\cdot)$, will be sufficient to impose a constraint of negativity (positivity) on the second-order finite differences. Again, this result can then be achieved by including an additional asymmetric penalty on the differences of second order, given by: $\sum_{k=1}^{K-2} I(\Delta^2 \delta_k > 0) \Delta^2 \delta_k$ for convexity or $I(\Delta^2 \delta_k \leq 0)$ for concavity.

If the k -th difference ($\Delta^2 \delta_k$) is less or equal to zero concavity is ensured and no penalty is required; otherwise, if $\Delta^2 \delta_k > 0$ it is necessary penalize such violation, as before, weighting it through a very high quantity, for instance $\kappa_2 = 10^6$. In matrix notation, the penalty becomes $\kappa_2 \boldsymbol{\delta}^T \mathbf{D}_2^T \mathbf{V}_2 \mathbf{D}_2 \boldsymbol{\delta}$ with $\mathbf{V}_2 = \text{diag}(I(\Delta^2 \delta_1 > 0), \dots, I(\Delta^2 \delta_{K-2} > 0))$, and the new objective function given by:

$$\| \mathbf{y} - \mathbf{B}\boldsymbol{\delta} \|^2 + \alpha \| \mathbf{D}_d \boldsymbol{\delta} \|^2 + \kappa_1 \| \mathbf{V}_1^{1/2} \mathbf{D}_1 \boldsymbol{\delta} \|^2 + \kappa_2 \| \mathbf{V}_2^{1/2} \mathbf{D}_2 \boldsymbol{\delta} \|^2;$$

with relevant estimates

$$\hat{\boldsymbol{\delta}}_\alpha = (\mathbf{B}^T \mathbf{B} + \alpha \mathbf{D}_d^T \mathbf{D}_d + \kappa_1 \mathbf{D}_1^T \mathbf{V}_1 \mathbf{D}_1 + \kappa_2 \mathbf{D}_2^T \mathbf{V}_2 \mathbf{D}_2)^{-1} \mathbf{B}^T \mathbf{y}.$$

Having defined the P-spline framework, it is necessary to consider an iterative procedure to get the parameters $\boldsymbol{\delta}$ satisfying the monotonicity and concavity constraints: starting from the estimates of the parameters $\boldsymbol{\delta}$ modeled with basic P-spline (i.e., no constraints), \mathbf{V}_1 and \mathbf{V}_2 are updated at each step to obtain new estimates of the parameters $\boldsymbol{\delta}$, until convergence is reached (i.e., $|\hat{\boldsymbol{\delta}}_{i+1} - \hat{\boldsymbol{\delta}}_i| < \epsilon$).

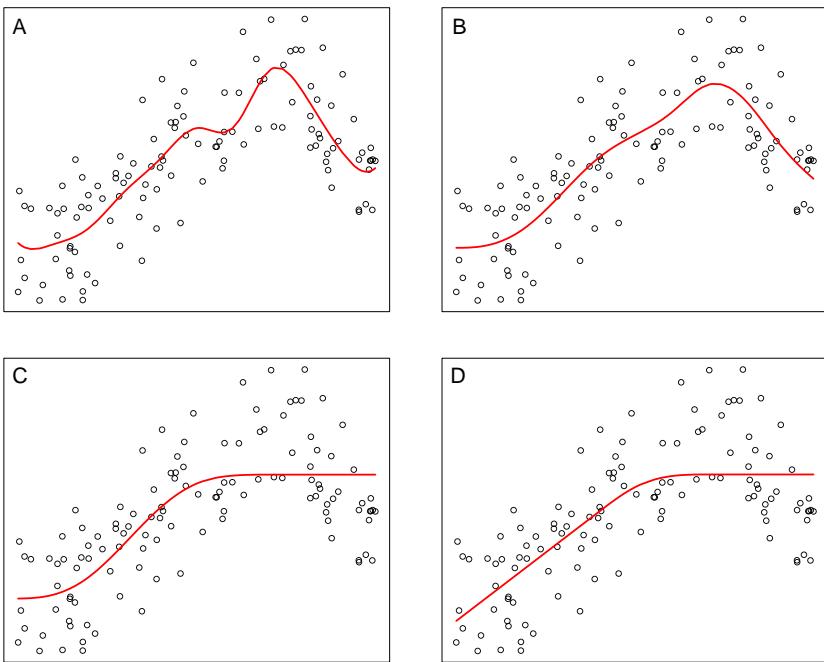


FIG. 1 B-spline (i), P-spline (ii), monotone P-spline (iii), and monotone and concave P-spline (iv).

Fig. 1 shows the potential of B-spline and P-spline (with and without constraint of monotonicity and concavity) in the evaluation of the cloud of points $\{x_i, y_i\}_{i=1}^n$.

3 Numerical illustrations

Ferrara and Vidoli (2017) consider four simulated data generating processes (DGP) in order to assess flexibility and robustness of the proposed GAM-SFA to departures from the standard stochastic frontier model. In all examples, the GAM-SFA fit is compared to the corresponding linear SFA with a Cobb–Douglas specification (i.e., on logs) thanks to the `semsfa` (Ferrara and Vidoli, 2018) and the `frontier` (Coelli and Henningsen, 2017) packages, respectively. All the R codes of the numerical examples are reported in Supplementary Materials available online at <https://doi.org/10.1016/bs.host.2018.11.004>.

3.1 Example 1: Linear homoscedastic model

In this first example, a linear homoscedastic data generating process is considered; more specifically, a sample of $n = 150$ observations has been generated from the following model

$$Y = 2 + 2 \times X + v - u,$$

where $X \sim U(0, 1)$, $u \sim |\mathcal{N}(0, \sigma_u = 1)|$, $v \sim \mathcal{N}(0, \sigma_v)$, and $\sigma_v = 0.75 \times \text{std}(U) = 0.75 \times \sigma_u \sqrt{(\pi - 2)/\pi}$. The estimated frontiers obtained for the single simulated sample from the GAM-SFA specification and the linear SFA (Fig. 2) are almost completely overlapping the real and unknown function: the linearity of the estimated models thus seems to suggest the choice of a linear production function; thin plate regression splines have been here adopted.

```

library(frontier)
library(semsfa)

set.seed(25)
n=150
x=runif(n,0,1)

u=abs(rnorm(n,0,1))
v=rnorm(n,0,.75*((pi-2)/pi))

y=2+2*x+v-u

dati<-data.frame(y,x)

#standard SFA
sfa<-sfa(y~x, data=dati)

#gAM-SFA
gamsfa<-semsfa(y~s(x,k=4,fx=TRUE,bs="tp"), data=dati)

```

3.2 Example 2: Nonlinear exponential homoscedastic model

Let us now consider a nonlinear exponential homoscedastic DGP

$$Y = 7 - 2 \times \exp(-X) + v - u,$$

where $X \sim U(0, 3)$, $u \sim |\mathcal{N}(0, \sigma_u = 1)|$, $v \sim \mathcal{N}(0, \sigma_v)$ and, as above, $\sigma_v = 0.75 \times \text{std}(U) = 0.75 \times \sigma_u \sqrt{(\pi - 2)/\pi}$. The graph in Fig. 3, based on a single drawn of $n = 150$ observations from the corresponding DGP, shows how the fitted GAM-SFA (with and without constraints) is able to capture the true curvature of the simulated production function, slightly better than the corresponding linear SFA with a Cobb–Douglas specification. It is easy to note that imposing shape constraints when they are true leads to small potential benefits and it may not outweigh the potential cost of misspecification. Thin plate regression splines have been here adopted for the unconstrained model and P-splines for constrained GAMs, respectively.

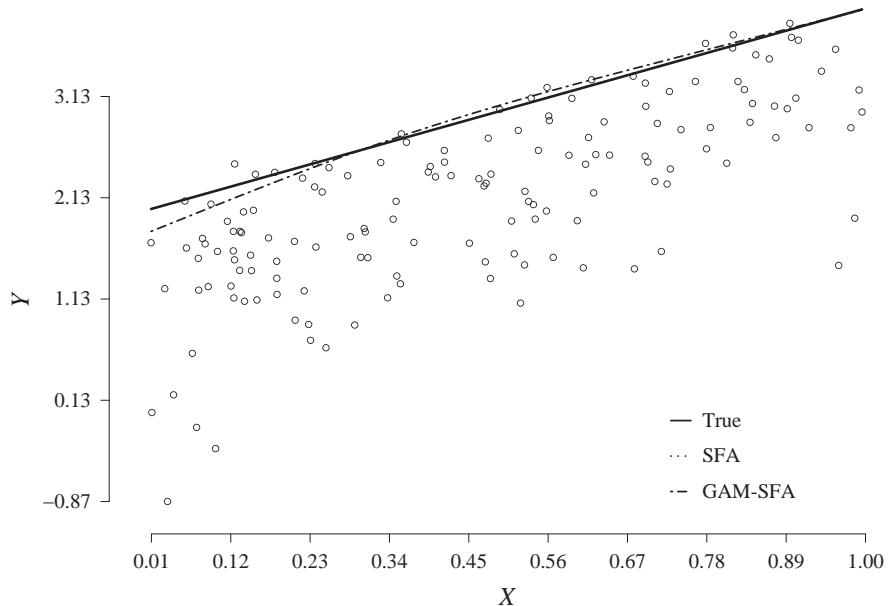


FIG. 2 Example 1—linear homoscedastic model: comparison of the true frontier (solid), the GAM-SFA fit (twodash), and the standard SFA (dotted).

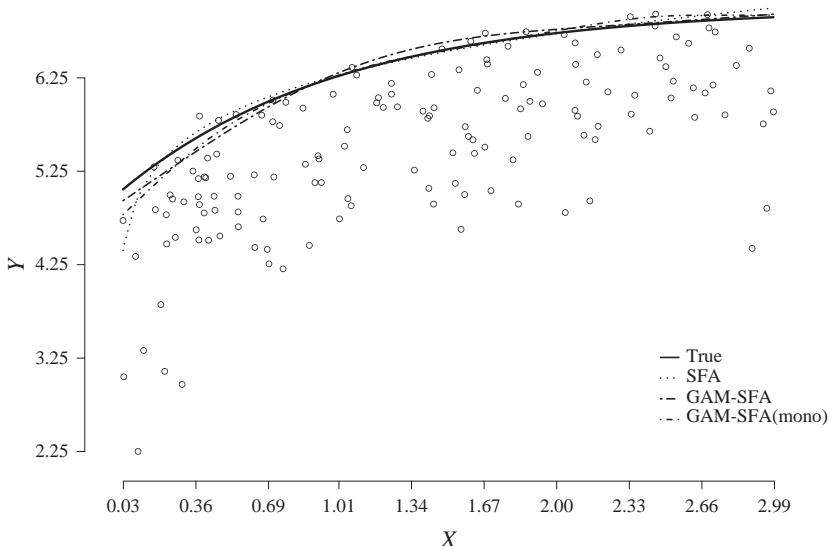


FIG. 3 Example 2—nonlinear exponential homoscedastic model: comparison of the true frontier (solid), the GAM-SFA fit (twodash), the monotone GAM-SFA fit (dotdash), and the standard SFA (dotted).

```

library(frontier)
library(semsfa)

set.seed(25)
n=150
x=runif(n,0,3)

u=abs(rnorm(n,0,1))
v=rnorm(n,0,.75*((pi-2)/pi))

y=7-2*exp(-x)+v-u

dati<-data.frame(y,x)

#standard SFA
sfa<-sfa(log(y)~log(x),dati)

#gamsfa
gamsfa<-semsfa(y~s(x,k=4,fx=TRUE,bs="tp"),dati)
gamsfa.monotone<-semsfa(y~pbm(x,mono="up"),sem.method
= "gam.monotone",data = dati)

```

3.3 Example 3: Nonmonotone model

Let us now consider a nonmonotone (first decreasing and then increasing) DGP, given the following sinusoidal signal:

$$Y = 3 \times \sin(2 \times X) + \text{uniform.error} + v - u,$$

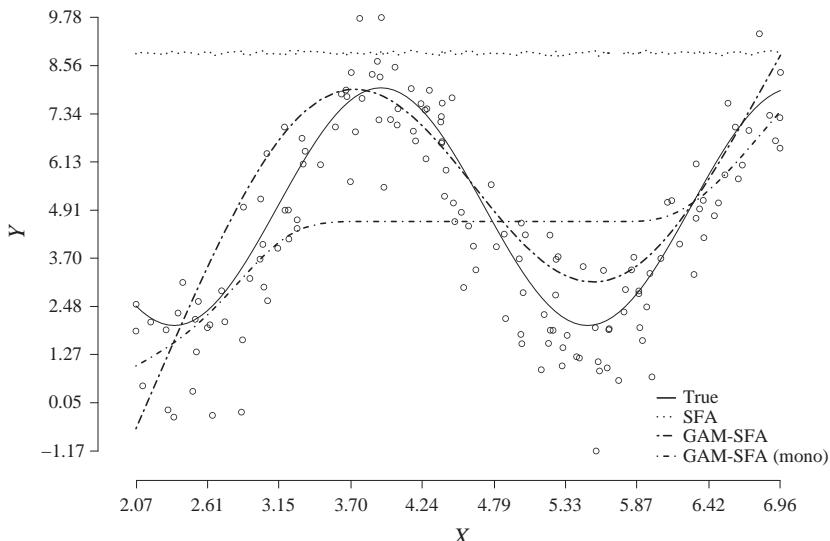


FIG. 4 Example 3—non monotone model: comparison of the true frontier (*solid*), the GAM-SFA fit (*twodash*), the monotone GAM-SFA fit (*dotdash*), and the standard SFA (*dotted*).

where $X \sim U(2, 7)$, $\text{uniform.error} = U(4, 7)$, $u \sim |\mathcal{N}(0, \sigma_u = 1)|$, $v \sim \mathcal{N}(0, \sigma_v)$ and, as above, $\sigma_v = 0.75 \times \text{std}(U) = 0.75 \times \sigma_u \sqrt{(\pi - 2)/\pi}$. The graph in Fig. 4, based on a single drawn of $n = 150$ observations from the corresponding DGP, shows how the fitted GAM-SFA without constraints is able to capture the true curvature of the simulated production function, in contrast to the corresponding constrained GAM-SFA and linear SFA (on logs).

```

library(frontier)
library(semsfa)

set.seed(1)
n <- 150
x <- runif(n, 2, 7)
uerr <- runif(n, 4, 7) # uniform error

u=abs(rnorm(n,0,1))
v=rnorm(n,0,.75*((pi-2)/pi))

y <- 3*sin(2*x) + uerr + v - u
dati<-data.frame(y,x)

#standard SFA
sfa<-sfa(log(y)~x,dati)

#gamsfa
gamsfa<-semsfa(y~s(x,k=4,fx=TRUE,bs="tp"),dati)
gamsfa.mono<-semsfa(y~pbm(x,mono="up"), data = dati,
                      sem.method = "gam.mono")

```

These results show that, when the true frontier is non monotone, the GAM-SFA without additional constraints is able to capture the relative true curvature and imposing shape constraints when they are not true leads to a potential cost of misspecification. Thin plate regression splines have been here adopted for the unconstrained model and P-splines for constrained GAMs, respectively.

3.4 Example 4: Quadratic polynomial model with heteroscedasticity

In the fourth scenario, the true DGP is not linear and it is characterized by heteroscedasticity of the inefficiency component u (i.e., σ_u is a function of the explanatory variable X):

$$Y = 5 + 25 \times X - 7 \times X^2 + v - u,$$

where $X \sim U(1, 2)$ and $u \sim |\mathcal{N}(0, \sigma_u = (1+X/2)^2)|$ and $v \sim \mathcal{N}(0, \sigma_v = 0.75)$.

The estimated frontier, as illustrated in Fig. 5 ($n = 200$), seems to follow the true curvature of the true model even if, as already mentioned, the implemented procedure does not handle explicitly the heteroscedasticity hypothesis; thin plate regression splines have been here adopted. This is a good evidence if compared with the standard SFA framework (Cobb–Douglas) when the homoscedasticity and the linearity assumptions fail simultaneously.

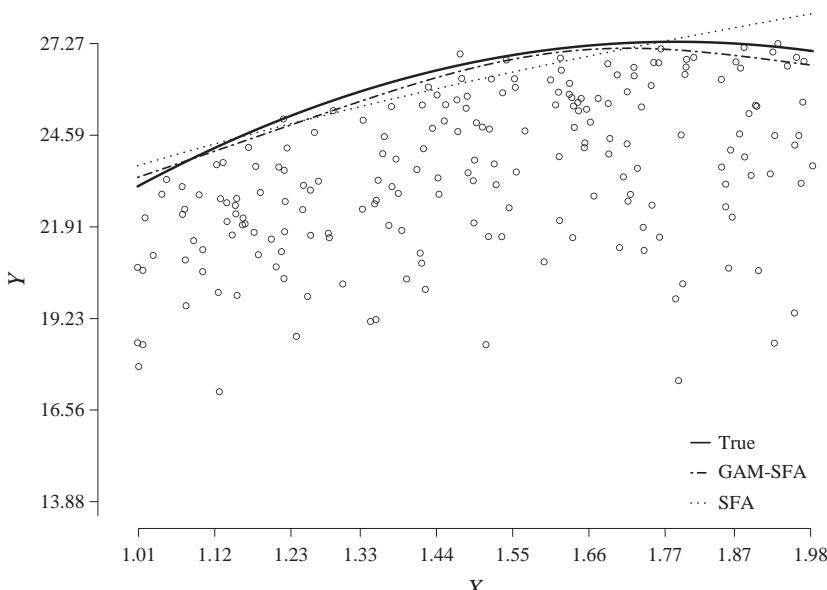


FIG. 5 Example 4—quadratic model with heteroscedasticity: comparison of the true frontier (solid), the GAM-SFA fit (twodash), and the standard SFA (dotted).

```
library(semsfa)
library(frontier)

set.seed(14)
n<-200
x<- runif(n, 1, 2)
fy<- exp(5+25*x-7*x^2)

v<- rnorm(n, 0, .75*((pi-2)/pi))
u<- abs(rnorm(n,0,(1+x/2)^2))

y <- log(fy) + v - u
dati<-data.frame(y,x)

#standard SFA
sfa<-sfa(log(y)~log(x),dati)

#GAM-SFA
gamsfa<-semsfa(y~s(x,k=4,fx=TRUE,bs="tp"),dati)
```

4 Empirical application to crops data

Over the last decades, the agricultural sector has undergone a significant structural change in EU countries; the expansion of average farm size and the decline in the number of farms are widely observed features, while several studies have been carried out on the farms productive efficiency. One of the first attempts to apply the efficiency theory and the practical application on the agricultural economy was made by [Coelli \(1995\)](#) recommending to use the stochastic frontier method because measurement error, missing variables and weather play a significant role in this field.

Here we consider the *Fieldcrops* production sector in the 2012 of the FADN data for the 109 main European regions in Supplementary Materials available online at <https://doi.org/10.1016/bs.host.2018.11.004>. The Farm Accountancy Data Network (FADN) is an annual survey carried out by the Member States of the European Union and it represents an instrument to evaluate the income of agricultural holdings and the impacts of the Common Agricultural Policy. Based on national surveys, FADN is the only source of microeconomic data that is harmonized, in other words, the bookkeeping principles are the same in all countries.

More specifically, it has been taken into account the *Total of output of crops and crop products* (SE135 code) as output and, as inputs, the following dimensions in Supplementary Materials available online at <https://doi.org/10.1016/bs.host.2018.11.004>: *Labor* expressed in AWU (annual work unit = full-time person equivalent—SE010 code), *Machinery* as a proxy of operating farm capital (machines, tractors, cars and lorries and irrigation equipment—SE455 code) and *Energy* as a proxy of the energy consumption (motor fuels and lubricants, electricity, heating fuels—SE345 code). The relationship between output and inputs has been initially analyzed according the standard linear SFA model which results are reported in [Table 1](#).

TABLE 1 Estimates for standard SFA

| | Estimate | Std. error | z value | $Pr(> z)$ |
|-----------|----------|------------|---------|--------------|
| Intercept | 2.560 | 0.490 | 5.223 | 1.760e-07*** |
| Labor | 0.007 | 0.056 | 0.128 | 0.898 |
| Energy | 0.535 | 0.083 | 6.403 | 1.522e-10*** |
| Machinery | 0.399 | 0.067 | 5.998 | 1.998e-09*** |

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

```

library(frontier)
library(semsfa)

# set your working directory (wd) for data and shape files
wd.name = "..."
setwd(wd.name)
# read data
datcsv = read.csv("dataFadn.csv")

# variables on log
datcsv$y= log(datcsv$SE135)
datcsv$x1=log(datcsv$SE010)#labor
datcsv$x2=log(datcsv$SE345)#energy
datcsv$x3=log(datcsv$SE455)#machinery

# standard sfa
prod.sfa <- sfa(y ~ x1 + x2 + x3, data = datcsv)
datcsv$eff.sfa<-as.vector(efficiencies(prod.sfa))#efficiency
summary(prod.sfa) # look at results

```

Labor it is clearly not statistically significant and the frontier model highlights the presence of inefficiency given that λ is equal to 2.003 (mean efficiency: 0.725).

Monotonicity hypothesis it does not necessarily hold in agricultural production processes where congestion (even locally) is considerable (Tone and Sahoo, 2004) given a limited availability of land, water, and other natural resources and given the presence of *bad outputs* (such as pollutants or other products causing disutility in the production process). For these reasons, the semiparametric approach appears a natural method for this kind of analysis. We can estimate our model with and without the assumption of monotonicity, considering the GAM-SFA specification with constrained and unconstrained P-splines.

```

# gam with monotonicity constraints
prod.gam.mono<-semsfa(y ~ pbm(x1,mono = "up")+
                         pbm(x2,mono = "up") + pbm(x3,mono = "up"),
                         data=datcsv, sem.method = "gam.mono",
                         ineffDecrease=TRUE)
datcsv$eff.gam.mono = efficiencies(prod.gam.mono)
$efficiencies

# to test significance in presence of monotonicity constraints
gam.full<-gamlss(y ~ pbm(x1,mono = "up")+
                     pbm(x2,mono = "up") + pbm(x3,mono = "up"),
                     data=datcsv)
gam.null<-gamlss(y ~ 1, data=datcsv)
# try fitting all models that differ from the current model
# by adding a single term
addterm(gam.null,gam.full,test = 'Chisq')

# gam without monotonicity constraints
prod.gam<-semsfa(y ~ ps(x1) + ps(x2) + ps(x3), data=datcsv
                  , sem.method = "gam.mono", ineffDecrease=TRUE)
datcsv$eff.gam = efficiencies.semsfa(prod.gam)$efficiencies
summary(prod.gam)

```

The GAM-SFA estimator provides efficiency estimates highly correlated with the corresponding SFA model (0.901) but, as indicated in [Table 2](#), it is possible to note that the *Labor* becomes more significant due to the more flexibility associated to the GAM framework; this finding is also highlighted by the λ estimate equals to 2.170.

For the model with monotonicity constrains, Likelihood ratio test (*LRT*) are used to evaluate the statistical significance of any nonlinear term. Estimation results are reported in [Table 3](#).

The GAM-SFA estimator with monotonicity constraints provides efficiency estimates highly correlated with the corresponding unconstrained GAM-SFA model (0.982), for this reason we deeply investigate efficiency

TABLE 2 Estimates for unconstrained GAM-SFA

| | Estimate | Std. error | t value | Pr(> t) |
|---------------|----------|------------|---------|--------------|
| ps(Labor) | 0.064 | 0.042 | 1.489 | 1.397e-01 |
| ps(Energy) | 0.528 | 0.069 | 7.692 | 1.335e-11*** |
| ps(Machinery) | 0.423 | 0.061 | 6.911 | 5.480e-10*** |

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

TABLE 3 Estimates for unconstrained GAM-SFA

| | LRT | Pr(Chi) |
|-----------|---------|-------------|
| Labor | 34.511 | 1.31e-13*** |
| Energy | 217.060 | <2.2e-16*** |
| Machinery | 191.697 | <2.2e-16*** |

* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

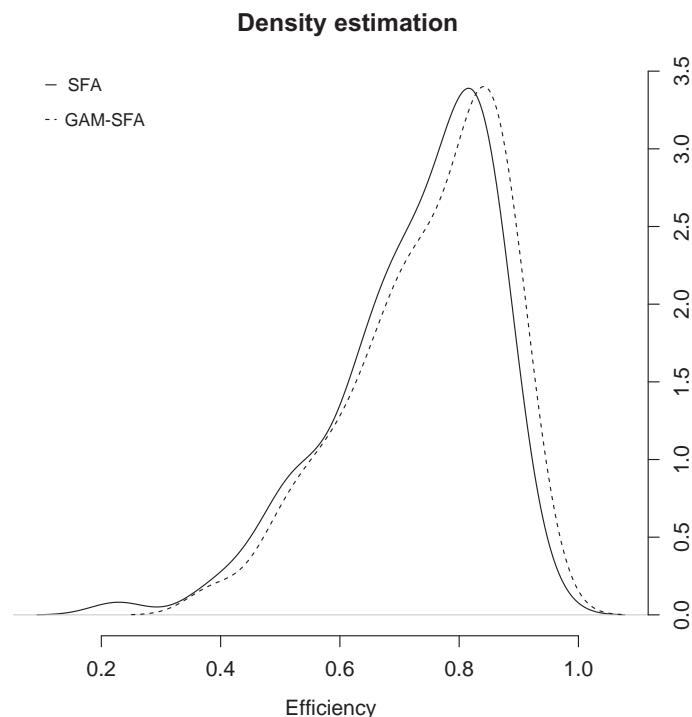


FIG. 6 Efficiency estimates from the standard SFA and the semiparametric unconstrained GAM-SFA, respectively.

estimates obtained from the standard SFA and the corresponding unconstrained GAM-SFA, respectively.

Fig. 6 reports density estimation of technical efficiencies obtained from SFA and unconstrained GAM-SFA, respectively; results appear quite similar for both models.

```

plot(density(datcsv$eff.sfa),type="l",lty=1,xlab=
    "efficiency", main="Density estimation",bty="n",
    yaxt="n",ylab="")
axis(4,cex=0.5)
points(density(datcsv$eff.gam),type="l",lty=2)
legend(x="topleft",c("sfa","gam-sfa"),bty="n",cex=.8,
    lty=c(1,2),seg.len=0.8)

```

Fig. 7 shows evident territorial patterns, again, similar for both specifications: Central European regions appear as the most efficient while the Northern ones, the Scotland and conversely the Southern ones as Greece show evidences of weakness regarding the crops production efficiency.

```

# read shapefile
library(rgdal)
shape = readOGR(dsn=wd.name, layer="shapeFadn")
# merge with estimates and data
shape@data = merge(shape@data,datcsv, by.x = "FADN_2012_",
    by.y = "FADN_2012_", all.x=TRUE)
library(lattice)
library(RColorBrewer)
trellis.par.set(axis.line=list(col=NA),
    strip.background=list(col=NA),
    strip.border=list(col=NA))
spplot(shape, c("eff.sfa","eff.gam"), layout=c(2,1),
    col.regions=colorRampPalette(brewer.pal(4,
    "Greys"))(18), col="grey", colorkey=list(space=
    "bottom"), names.attr=c("SFA", "GAM-SFA"))

```

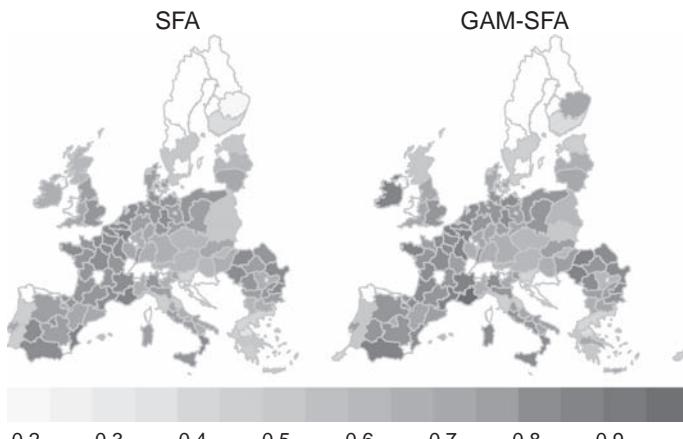


FIG. 7 Efficiency territorial distribution related to the standard SFA and the semiparametric unconstrained GAM-SFA, respectively.

5 Conclusions

A better understanding of the relationship among inputs and output is relevant to properly adopt specific policies in order to avoid a condition of systematic inefficiency.

Since the true form of the frontier is typically unknown and consequently it is impossible to measure how well any chosen form approximates the true, the concept of flexibility is extremely relevant to make a more accurate model selection; alternative functional forms and comparisons of alternatives should be the best approach for applied and empirical analyses. Indeed, restricting the focus to the most frequently functional forms can lead to misleading conclusions. For this reason, we have analyzed possible different specification of stochastic frontier models starting from the packages already available in the R environment.

The capability of the proposed GAM framework for the analysis of production frontier has been illustrated by some numerical illustrations and one empirical analysis; the approach seems to be very useful for efficiency analysis especially if there is a lack of knowledge about the specification of the functional form of the underlying process. Allowing nonlinear dependence between inputs and the output, the GAM specification is likely to improve the overall fit of the model as compared to the fully parametric specification.

As highlighted by [Henningsen and Kumbhakar \(2009\)](#), although in many empirical applications [Fan et al. \(1996\)](#) approach seems to be more appropriate than the full parametric SFA and DEA, it has not been widely used in applied studies, probably because of nonavailability of user-friendly software. For this reason the `semsfa` R package has been introduced trying to overcome the nowadays mentioned limit; future releases will allow to apply all the proposed methods.

References

- Adetutu, M., Glass, A., Kenjegalieva, K., Sickles, R., 2015. The effects of efficiency and TFP growth on pollution in Europe: a multistage spatial analysis. *J. Prod. Anal.* 43 (3), 307–326.
- Aigner, D.J., Chu, S.F., 1968. On estimating the industry production function. *Am. Econ. Rev.* 58 (4), 826–839.
- Aigner, D., Lovell, C., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production function models. *J. Econ.* 6 (1), 21–37.
- Anselin, L., 2001. Spatial effects in econometric practice in environmental and resource economics. *Am. J. Agric. Econ.* 83 (3), 705–710.
- Areal, F.J., Balcombe, K., Tiffin, R., 2012. Integrating spatial dependence into stochastic frontier analysis. *Aust. J. Agric. Resour. Econ.* 56, 521–541.
- Battese, G.E., Coelli, T., 1995. A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empir. Econ.* 20 (2), 325–332.
- Billé, A.G., Salvioni, C., Benedetti, R., 2018. Modelling spatial regimes in farms technologies. *J. Prod. Anal.* 49 (2), 173–185.
- Bollaerts, K., Eilers, P. H. C., Aerts, M., 2006. Quantile regression with monotonicity restrictions using P-splines and the L1-norm. *Stat. Model.* 6 (3), 189–207.

- Cleveland, W.S., Grosse, E., Shyu, W.M., 1993. Local regression models. In: Chambers, J.M., Hastie, T.J. (Eds.), *Statistical Models* in S. Chapman & Hall, New York.
- Coelli, T., 1995. Recent developments in frontier modelling and efficiency measurement. *Aust. J. Agric. Econ.* 39 (3), 219–245.
- Coelli, T., Henningsen, A., 2017. *frontier: Stochastic Frontier Analysis*. R package version 1.1-2. <https://CRAN.R-Project.org/package=frontier>.
- Currie, I.D., Durban, M., 2002. Flexible smoothing with P-splines: a unified approach. *Stat. Model.* 2 (4), 333–349.
- de Boor, C., 1978. *A Practical Guide to Splines*. Springer.
- Druska, V., Horrace, W., 2004. Generalized moments estimation for spatial panel data: indonesian rice farming. *Am. J. Agric. Econ.* 86 (1), 185–198.
- Eilers, P. H. C., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. *Stat. Sci.* 11, 89–121.
- Fan, Y., Li, Q., Weersink, A., 1996. Semiparametric estimation of stochastic production frontier models. *J. Bus. Econ. Stat.* 14 (4), 460–468.
- Farrell, M.J., 1957. The measurement of productive efficiency. *J. R. Stat. Soc. Ser. A (General)* 120 (3), 253–290.
- Ferrara, G., Vidoli, F., 2017. Semiparametric stochastic frontier models: a generalized additive model approach. *Eur. J. Oper. Res.* 258 (2), 761–777.
- Ferrara, G., Vidoli, F., 2018. *semsfa: semiparametric estimation of stochastic frontier models*. R package version 1.1. <https://CRAN.R-Project.org/package=semsfa>.
- Florens, J.-P., Simar, L., Keilegom, I., 2014. Frontier estimation in nonparametric location-scale models. *J. Econ.* 178, 456–470.
- Fusco, E., Vidoli, F., 2015. *ssfa: spatial stochastic frontier analysis*. R package version 1.1. <https://CRAN.R-project.org/package=ssfa>.
- Giannakas, K., Tran, K., Tzouvelekis, V., 2003. On the choice of functional form in stochastic frontier modeling. *Empir. Econ.* 28, 75–100.
- Glass, A., Kenjegalieva, K., Sickles, R., 2014. Estimating efficiency spillovers with state level evidence for manufacturing in the US. *Econ. Lett.* 123, 154–159.
- Glass, A.J., Kenjegalieva, K., Sickles, R.C., 2016. A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillovers. *J. Econ.* 190 (2), 289–300.
- Greene, W., 2008. The econometric approach to efficiency analysis. In: Fried, H.O., Knox Lovell, C.A., Schmidt, S.S. (Eds.), *The Measurement of Productive Efficiency and Productivity Change*. Oxford University Press.
- Hastie, T.J., Tibshirani, R.J., 1990. *Generalized Additive Models*. Chapman & Hall, London.
- Hayfield, T., Racine, J.S., 2008. Nonparametric econometrics: the np package. *J. Stat. Softw.* 27(5). <http://www.jstatsoft.org/v27/i05/>.
- Henningsen, A., Kumbhakar, S., 2009. Semiparametric stochastic frontier analysis: an application to polish farms during transition. European Workshop on Efficiency and Productivity Analysis. EWEPA, Pisa, Italy.
- Johnson, A., Kuosmanen, T., 2011. One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. *J. Prod. Anal.* 36 (2), 219–230.
- Jondrow, J., Lovell, C., Materov, I.S., Schmidt, P., 1982. On the estimation of technical inefficiency in the stochastic frontier production function model. *J. Econ.* 19 (2–3), 233–238.
- Kumbhakar, S.C., Lovell, C. A. K., 2000. *Stochastic Frontier Analysis*. Cambridge University Press.

- Kumbhakar, S., Ghosh, S., McGuckin, J.T., 1991. A generalized production frontier approach for estimating determinants of inefficiency in U.S. dairy farms. *J. Bus. Econ. Stat.* 9 (3), 279–286.
- Kumbhakar, S., Park, B.U., Simar, L., Tsionas, M., 2007. Nonparametric stochastic frontiers: a local maximum likelihood approach. *J. Econ.* 137 (1), 1–27.
- Kuosmanen, T., 2008. Representation theorem for convex nonparametric least squares. *Econ. J.* 11 (2), 308–325.
- Kuosmanen, T., Johnson, A., 2010. Data envelopment analysis as nonparametric least-squares regression. *Oper. Res.* 58 (1), 149–160.
- Kuosmanen, T., Kortelainen, M., 2012. Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. *J. Prod. Anal.* 38 (1), 11–28.
- Latruffe, L., 2010. Competitiveness, productivity and efficiency in the agricultural and agri-food sectors. OECD Publishing. OECD Food, Agriculture and Fisheries Working Papers, No. 30.
- Lee, C.-Y., Johnson, A.L., Moreno-Centeno, E., Kuosmanen, T., 2013. A more efficient algorithm for convex nonparametric least squares. *Eur. J. Oper. Res.* 227 (2), 391–400.
- Lensink, R., Meesters, A., 2014. Institutions and bank performance: a stochastic frontier analysis. *Oxf. Bull. Econ. Stat.* 76 (1), 67–92.
- Mastromarco, C., Simar, L., 2015. Effect of FDI and time on catching up: new insights from a conditional nonparametric frontier analysis. *J. Appl. Econ.* 30 (5), 826–847.
- McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, seconnd ed. Chapman & Hall.
- Meeusen, W., van den Broeck, J., 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. *Int. Econ. Rev.* 18 (2), 435–444.
- Muggeo, V. M. R., Ferrara, G., 2008. Fitting generalized linear models with unspecified link function: a P-spline approach. *Comput. Stat. Data Anal.* 52 (5), 2529–2537.
- Pavlyuk, D., 2016. spfrontier: spatial stochastic frontier models. R package version 0.2.3. <https://CRAN.R-project.org/package=spfrontier>.
- R Development Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org>.
- Racine, J., Su, L., Ullah, A., 2014. The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics. Oxford University Press.
- Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C* 54 (3), 507–554.
- Schmidt, A., Moreira, A., Helfand, S., Fonseca, T., 2009. Spatial stochastic frontier models: accounting for unobserved local determinants of inefficiency. *J. Prod. Anal.* 31 (2), 101–112.
- Simar, L., Wilson, P., 2008. Stochastic panel frontiers: a semiparametric approach. In: Fried, H.O., Knox Lovell, C.A., Schmidt, S.S. (Eds.), *The Measurement of Productive Efficiency and Productivity Change*. Oxford University Press.
- Stasinopoulos, D., Rigby, R., 2007. Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Softw.* 23 (7), 1–46.
- Stasinopoulos, D., Rigby, R., Akantziliotou, C., 2006. Instructions on how to use the gamlss package in R. London Metropolitan University. Technical report, STORM Research Centre.
- Stone, C., 1986. The dimensionality reduction principle for generalized additive models. *Ann. Stat.* 14, 590–606.
- Tone, K., Sahoo, B., 2004. Degree of scale economies and congestion: a unified DEA approach. *Eur. J. Oper. Res.* 158 (3), 755–772.
- Tsionas, M., Michaelides, P., 2016. A spatial stochastic frontier model with spillovers: evidence for Italian regions. *Scottish J. Polit. Econ.* 63 (3), 243–257.

- Vidoli, F., Ferrara, G., 2015. Analyzing Italian citrus sector by semi-nonparametric frontier efficiency models. *Empir. Econ.* 49 (2), 641–658.
- Wood, S.N., 2003. Thin plate regression splines. *J. R. Stat. Soc. Ser. B* 65 (1), 95–114.
- Wood, S.N., 2017. An Introduction to Generalized Additive Models With R, second ed. Chapman and Hall/CRC.
- Wood, S.N., 2018. mgcv: mixed GAM computation vehicle with automatic smoothness estimation. R package version 1.8-24. <https://CRAN.R-project.org/package=mgcv>.

Index

Note: Page numbers followed by “*f*” indicate figures, “*t*” indicate tables, “*b*” indicate boxes, and “*np*” indicate footnotes.

A

- Activation function, 97–101
- Activity analysis models (AAM), 280–281
- AdaBoost, 83–90
- ada* package, 86
- ADL-MIDAS, 207–209
- Aït-Sahalia and Jacod (ASJ) test, 5, 16–17
- Akaike information criterion (AIC), 197–198, 248, 310

Algorithm

- Discrete AdaBoost, 84*b*
- Gentle AdaBoost, 94*b*
- LogitBoost, 92–93, 92*b*
- Real AdaBoost (RAB), 91*b*
- SIM-Rodeo, 107*b*
- Astsa (Applied Statistical Time Series Analysis), 216–217
- Augmented Dickey–Fuller (ADF) test, 62, 64–65, 194–195
- Autocorrelation functions (ACF), GDP and GDP deflator, 195
- Autoregressive distributed lag (ARDL) models, 158
- Autoregressive integrated moving average (ARIMA) model, 195–197 and MIDAS, 213–214
- Average inefficiency, 272

B

- BAND-TAR, 232
- Barndorff-Nielsen and Shephard (BNS) test, 5, 15
- Bayesian information criteria (BIC), 194–195, 248
- Bernoulli distribution, 85–86, 107–108
- Bernoulli log-likelihood, 92–93
- Bipower variation (BPV), 4
- BLT co-jump testing, 20–21
- Bootstrap, 72, 156–157, 159, 236, 289
- Box–Jenkins univariate time-series models, 196–197
- Bridge equations, 188, 201–202 using principal components, 203–205
- Brownian motion, 6–8
- Brownian semimartingale process, 6–8
- B-splines, 300, 309–310, 312*f*
- Bubble identification, PSY procedure, 68–70 rationale, 68
- Bureau of Economic Analysis (BEA), 201

C

- caTools*, 93–94
- Causality, 164, 173, 175–177, 175*t*
- CCR model, 280
- Circle_data* function, 85–86, 86*b*
- Circle model, 108
- Cobb–Douglas production function, 270–272
- Co-exceedance rule, 19–20, 22–23
- Cointegration, 230–231 estimation and testing for, 233–234 residual-based approach, 233–234
- Co-jump, 5
- Co-jump tests, 4–5
 - BLT, 20–21
 - CKR, 23–24
 - GST co-exceedance rule, 22–23
 - Jacod and Todorov, 21–22
 - Mancini and Gobbi threshold, 22
- Compound Poisson process (CPP), 6–8
- Consumer price index (CPI), 162
- Consumer price inflation (CPI), 176
- Contextual variables, 305–307
- Corradi, Silvapulle, and Swanson (CSS) test, 18–19
- Corrected ordinary least squares, 271
- Correlations and cross-correlations, GDP and GDP deflator, 195–196
- CPI. *See* Consumer price index (CPI)
- Credit risk, in European sovereign sector, 75–77
- Crisis identification, PSY procedure consistency, 71 rationale, 70–71

Crisis periods

- in European sovereign sector, 76, 76^f
- in S&P 500 stock market, 73–74, 73^f

Crowding out

- effect of government investment during 1970–2013, 157–158
- effects of public investment, 172
- effects on private investment, 157, 158^f

Current Quarterly Model (CQM), 188

D

- DAB. *See* Discrete AdaBoost (DAB)
- Data envelopment analysis (DEA), 267–268, 280, 301
- DEA-CRS model, 283–284
 - DEA-NIRS (nonincreasing returns to scale), 284
 - DEA-VRS (variable returns to scale), 284
 - envelopment form of, 281
 - model formulation, 280–283
 - origins, 280
 - output-oriented Farrell technical efficiency, 285
 - R codes, 289–290
 - statistical analysis, 286–287
 - streams of, 285–286
- Data generating process (DGP), 126
- Decision-making unit (DMU), 280, 308
- Deep Neural Network (DNN), 96–102, 96^f, 110–111
- Difference-stationary, 230
- Directional distance function, 282–283
- Direct test, threshold cointegration model, 239
- Discrete AdaBoost (DAB), 82, 84^b, 87–90
- DNN. *See* Deep Neural Network (DNN)
- DOW, 5
- Dynamic factor models indicators
- large number of, 219–222
 - small number of, 216–219
- Dynamic linear models (DLM), 216–217

E

- Economic surveys, India, 156
- Error correction term (ECT), 232
- ETFs, 5–6
- European sovereign debt sector, PSY detection, 75–77
- Exponential loss, 87–89
- Exxon Mobile Corporation (XOM), 5

F

- Farm Accountancy Data Network (FADN), 318
- Farrell's approach, 280
- fastAdaboost*, 87
- Fast Kalman Filter (FKF), 216–217
- Federal funds rate, inflation and effective, 149–150, 151^f, 151^t
- Fieldcrops production sector, 318
- Financial econometrics
- co-jump testing
 - BLT, 20–21
 - CKR, 23–24
 - GST co-exceedance rule, 22–23
 - Jacod and Todorov, 21–22
 - Mancini and Gobbi threshold, 22
- empirical experiments
- data description, 24–25
 - findings, 26–48
 - methodology, 25–26
- integrated volatility, 8–14
- MedRV*, 14
 - MinRV*, 14
 - modulated bipower variation, 12
 - multiscale realized volatility, 11
 - realized bipower variation, 9
 - realized kernel, 11
 - realized volatility, 9
 - subsampled realized kernel, 13
 - threshold bipower variation, 12–13
 - tripower variation, 10
 - truncated realized volatility, 12
 - two-scale realized volatility, 10–11
- jump testing, 14–19
- Ait-Sahalia and Jacod test, 16–17
 - Barndorff-Nielsen and Shephard test, 15
 - Corradi, Silvapulle, and Swanson test, 18–19
 - Jiang and Oomen test, 16–17
 - Lee and Mykland test, 15–16
 - Podolskij and Ziggel test, 18
- Firm, 268^{np}
- Fiscal policy, 164–166
- Forecasting models
- MIDAS regressions, 189–190
 - mixed-frequency dynamic latent factor models, 190–191
 - monthly models, 189
 - quarterly models, 188
- Free disposal hull (FDH), 284
- statistical analysis, 286–287

G

GAB. *See* Gentle AdaBoost (GAB)
 GAMLSS. *See* Generalized additive models
 for location, scale, and shape
 (GAMLSS)
 gamlss package, 306
 GAM-SFA
 linear homoscedastic model, 312–313, 314f
 nonlinear exponential homoscedastic
 model, 313–315, 315f
 nonmonotone model, 315–317, 316f
 quadratic polynomial model with
 heteroscedasticity, 317–318, 317f
 Gaussian kernel, 105–106
 Generalized additive models (GAM), 300, 302
 Generalized additive models for location,
 scale, and shape (GAMLSS), 305–306
 Generalized boosting regression models
 (GBM), 87
 Generalized impulse response functions
 (GIRF), 240–243
 Gentle AdaBoost (GAB), 94–96, 94b
glmnet package, 102–103
 Granger causality test, 197–198
 GST co-exceedance rule, 22–23

H

Hannan–Quinn criterion, 197–198, 248
 Hausman–Taylor estimators, 275–276,
 288–289
 Heteroskedasticity, 63, 72
 High-dimensional binary classification and
 probability prediction
 AdaBoost, 83–90
 Deep Neural Network, 96–102, 96f
 Discrete AdaBoost, 82, 84b, 87–90
 Gentle AdaBoost, 94–96, 94b
 logistic regression, 102–104
 Real AdaBoost, 91–92, 91b
 SIM-RODEO, 82–83, 105
 High-dimension (sparse) circle model, error
 rate of, 108, 109t
 High-dimension (sparse) logistic model, error
 rate of, 109–110, 110t
 Hodrick–Prescott filter (HP filter), 164

I

Impulse response function (IRF), 240–243,
 241–242f
 Index model. *See* Single index MIDAS model
 (SI-MIDAS)

Infrastructure investment, 157, 165f, 166,
 172–173, 177–178
 Input oriented DEA, 281
 Integrated volatility, 4, 6–14
 MedRV, 14
 MinRV, 14
 modulated bipower variation, 12
 multiscale realized volatility, 11
 realized bipower variation, 9
 realized kernel, 11
 realized volatility, 9
 subsampled realized kernel, 13
 threshold bipower variation, 12–13
 tripower variation, 10
 truncated realized volatility, 12
 two-scale realized volatility, 10–11

J

Jarque–Bera test, 198
 Jeon and Sickles (JS) R codes, 289
 Jiang and Oomen test, 16–17
 Johansen trace test, 248
JOUSBoost, 85–86
 JPMorgan Chase & Co. (JPM), 5
 JT co-jump testing, 21–22
 Jump, 4–5. *See also* Co-jump
 Jump diffusion, 14–15
 Jump tests, 4–5, 14–19
 Aït-Sahalia and Jacod test, 16–17
 Barndorff-Nielsen and Shephard test, 15
 Corradi, Silvapulle, and Swanson test, 18–19
 Jiang and Oomen test, 16–17
 Lee and Mykland test, 15–16
 Podolskij and Ziggel test, 18

K

Kalman filter, 119, 164, 191
keras, 98–102
 Kernel estimator, 104–105
 KSS model, 278

L

Least absolute shrinkage and selection operator
 (LASSO), 82–83, 118
 Lee and Mykland (LM) test, 5, 15–16
 Likelihood ratio test (LRT), 320
 Linear and quasi-linear MIDAS models
 (affine g), 123–128
 nonparametric smoothing of weights, 127–128
 unconstrained MIDAS, 124–125
 Linear homoscedastic model, 312–313, 314f

Linear models, tsDyn package, 243–244
 Logistic Regression, 111
 with LASSO, 102–104
 Logistic smooth transition MIDAS
 (LSTR-MIDAS), 129–130
 Logit, 104–105
 LogitBoost (LB), 92–93, 92*b*, 109–110
 Log-price, 6–8
 Low-dimension circle model, error rate of,
 108, 108*t*
 Low-dimension logistic model, error rate of,
 109–110, 110*t*
 LSTR-MIDAS. *See* Logistic smooth transition
 MIDAS (LSTR-MIDAS)

M

Mam濶quist indexes, 289
 Market crashes, 70
 Matconv package, 290
 Maximum entropy (ME) bootstrap method,
 156–157
 McKinnon hypothesis, 161, 177
 Meboot algorithm, 159–160
MedRV, 14
 MG threshold co-jump test, 22
 Microsoft Corporation (MSFT), 5
 MIDAS. *See* Mixed data sampling (MIDAS)
 MIDAS-VAR, 209–211
 MIDAS with min–mean–max effects
 (MMM-MIDAS), 130–131
 MIDAS with partially (quasi)linear effects
 (PL-MIDAS), 133–134
MinRV, 14
 Mixed data sampling (MIDAS) model, 205–207
 autoregressive integrated moving average
 model and, 213–214
 factor, 214–216
 regressions, 189–190
 Mixed data sampling (MIDAS) regression,
 118–119
 Mixed data sampling (MIDAS) regression
 model
 illustration with simulated data, 135
 data generation, 135–137
 estimation, 137–144, 140*t*, 141–143*f*
 linear and quasi-linear MIDAS models
 (affine g), 123–128
 nonparametric smoothing of
 weights, 127–128
 unconstrained MIDAS, 124–125
 nonlinear parametric MIDAS models,
 128–129

logistic smooth transition MIDAS, 129–130
 MIDAS with min–mean–max
 effects, 130–131
 semiparametric MIDAS models, 131–135
 MIDAS with partially (quasi)linear
 effects, 133–134
 single index MIDAS model, 134–135,
 149, 151*f*
 stylized regression model, 119–123
 constraint function h , 121
 Mixed frequency data, 119, 144–145
 Mixed-frequency dynamic latent factor models
 (MF-DLFM), 190–191
 Mixed-Frequency Vector Autoregressive
 (MF-VAR), 189
 Modulated bipower variation (MBV), 12
 Monetary policy, 156, 164–166
 Monotonicity, 309–311, 319–320
 Monte Carlo simulations, 22, 107–111
 Monthly models, 189
 Multiplicity/family-wise size control, 63
 Multiscale realized volatility (MSRV), 11
 Multivariate Autoregressive State-Space
 Modeling (MARSS), 216–217
 Multivariate models, tsDyn package, 245–246
 Multivariate system-based approach, threshold
 cointegration in, 256–260

N

National Income and Product Accounts
 (NIPA), 201
 Nearest neighbor truncation, 8–9, 14
neuralnet, 98–101
 Nodes, 309–310
 Nonlinear exponential homoscedastic
 model, 313–315, 315*f*
 Nonlinear parametric MIDAS models, 128–129
 logistic smooth transition MIDAS, 129–130
 MIDAS with min–mean–max effects,
 130–131
 Nonmonotone model, 315–317, 316*f*
 Nonparametric stochastic frontiers, 279–280, 287

O

Okun’s law, 145–149, 146*t*, 148*f*
 One-variable logistic regression, 85
 Ordinary least squares (OLS), 124

P

Panel stochastic production frontiers, 274–276
 Parametric random effects model, 274–275

- Partial autocorrelation functions (PACF),
GDP and GDP deflator, 195
- Philippine GDP and inflation
forecasting models
MIDAS regressions, 189–190
mixed-frequency dynamic latent factor
models, 190–191
monthly models, 189
quarterly models, 188
- getting data with R, 191–192
indicators, 191
- quarterly real GDP and GDP deflator
ADL-MIDAS, 207–209
ARIMA and MIDAS, 213–214
autocorrelation and partial
autocorrelation, 195
- Box–Jenkins univariate time-series
models, 196–197
- bridge equations, 201–202
- correlations and cross-correlations,
195–196
- descriptive statistics, 193–194
- dynamic factor models, 216–222
- factor MIDAS, 214–216
- indicators, 191–192
- MIDAS models, 205–207
- MIDAS-VAR, 209–211
- principal components, 203–205
- unit root tests for, 194–195
- VARX, 211–213
- VARXM, 211–213
- vector autoregressive models, 197–201
- PL-MIDAS. *See* MIDAS with partially
(quasi)linear effects (PL-MIDAS)
- Podolskij and Ziggel (PZ) test, 18
- Price-dividend ratios, 72–73
of S&P 500 index, 73–74, 73*f*
- Principal components, using monthly
indicators, 203–205
- Private investment, India, 155–156
causality results, 175–177, 175–176*t*
crowding out effects on, 157, 158*t*
data abbreviations and sources, 173–174, 174*t*
estimation, 164–173, 167–171*t*
interpreting data and model implications,
161–164, 162*t*, 163*f*, 165*f*
- Probit, 104–105
- Production efficiency, 268–269
- Production function, 270–272
- P-splines, 300, 304, 308–312
- Psych package, 194
- psymonitor R package, 63–64, 67
- PSY procedure
- augmented Dickey–Fuller test, 64–65
- bubble identification
consistency, 68–70
rationale, 68
- crisis identification
consistency, 71
rationale, 70–71
- European sovereign debt sector, 75–77
- potential of, 62
- recursive evolving algorithm, 66–68, 66*f*
- S&P 500 stock market, 72–75
- ## Q
- Quadratic polynomial model with
heteroscedasticity, 317–318, 317*f*
- Quarterly models, 188
- ## R
- RAB. *See* Real AdaBoost (RAB)
- R code, 49–56, 276–277
for DEA models, 289–290
for SFA models, 287–289
- Real AdaBoost (RAB), 91–92, 91*b*, 109–110
- Real interest rate, 156, 161–162, 164–166, 176
- Realized bipower variation (BPV), 9
- Realized kernel (RK), 11
- Realized quadratic variation, 15
- Realized volatility (RV), 4, 8–9
- Real-time bubble detection method. *See* PSY
procedure
- Rectified Linear Unit (ReLU), 97
- Recursive evolving algorithm, 63, 66–68, 66*f*
- Regularization of the Derivative Expectation
Operator (RODEO), 82–83
- Residual-based approach, cointegration, 233–234
- R functions, 98–101
in economics applications, 111, 111*t*
- Ridge regression, 118
- Russell efficiency measure, 282
- ## S
- Schwarz information criterion, 197–198
- Self-exciting threshold autoregressive model
(SETAR), 231–232, 234, 244–245,
251–252
- Semiparametric MIDAS models
MIDAS with partially (quasi)linear effects,
133–134
single index MIDAS model, 134–135, 151*f*
- Semiparametric regression, 119
- Semiparametric single-index model, 104–107
- semsfa package, 304, 304*np*

SETAR. *See* Self-exciting threshold autoregressive model (SETAR)
 Shephard output distance function, 270
 Shephard's distance functions, 282
 SI-MIDAS. *See* Single index MIDAS model (SI-MIDAS)
 SIM-RODEO, 82–83, 105
 Single index MIDAS model (SI-MIDAS), 134–135, 149, 151^f
 Single-index model, semiparametric, 104–107
 Smoothed randomized realized variance (SRRV), 23
 Smoothed realized variance (SRV), 23
 Smooth transition autoregressive (STAR) model, 243
 Smooth transition models, 129–130
 Spatial dependence, 307–308
 Spatial heterogeneity, 307
 Spatial lag, 308
 spfrontier package, 308
 S&P 500 stock market, PSY detection, 72–75
 ssfa package, 308
 SS models. *See* State space (SS) models
 SSR, 235
 State space (SS) models, 119
 Stochastic frontier analysis (SFA), 267–268, 270–271
 contextual variables, 305–307
 corrected ordinary least squares, 271
 efficiency estimation, 273–274
 empirical application to crops data, 318–322
 and factor models, 277–278
 methods, 301–312
 numerical illustrations
 linear homoscedastic model, 312–313
 nonlinear exponential homoscedastic model, 313–315
 nonmonotone model, 315–317
 quadratic polynomial model with heteroscedasticity, 317–318
 panel data model, 274–276
 P-splines, 308–312, 312^f
 R codes, 287–289
 spatial external factors, 307–308
 true fixed effects model, 278–280
 StoNED framework, 304
 Structural break tests, 236
 Stylized MIDAS regression model, 119–123
 constraint function h , 121

selection of h , d , and k , 121–122
 statistical inference, 122–123
 Subsampled realized kernel (SRK), 13
 Sup-LM test, 238
 Swap variance approach, 4–5, 14–15

T

Technical efficiency. *See* Production efficiency
 Threshold bipower variation (TBPV), 8–9, 12–13
 Threshold cointegration model
 estimation and testing for, 237–239, 238^t
 direct test, 239
 remarks, 239
 two-step approach, 238
 generalized impulse response functions, 240–243
 impulse response function, 240–243,
 241–242^f
 joint estimation approach, 240
 linear and, 230–233
 in multivariate system based approach,
 256–260
 tsDyn package
 linear models, 243–244
 multivariate models, 245–246
 univariate models, 244–245
 two-step approach, 240
 unit roots and linear cointegration, 247–250
 in univariate residual-based approach,
 250–256

Threshold estimated model
 estimation and testing for, 234–236
 estimation of, 240
 Threshold VECM (TVECM), 232
 Time series, 156–161, 164, 166, 172, 177
 Tripower variation (TPV), 4, 10
 True effects model, 276–280
 Truncated realized volatility (TRV), 12
 tsDyn package, threshold cointegration model
 linear models, 243–244
 multivariate models, 245–246
 univariate models, 244–245
 Two-scale realized volatility (TSRV), 10–11

U

Unconstrained MIDAS model, 118, 125–126, 142
 Unit root tests
 for GDP and GDP deflator, 194–195

and linear cointegration, 247–250
Univariate model, tsDyn package, 244–245
Univariate residual-based approach, threshold cointegration, 250–256

V

VARXM model, 211–213
VARX model, 211–213
Vector autoregressive (VAR) model, 158, 197–201, 230, 243–244
Vector error correction model (VECM), 230, 239, 243–244

W

Walmart Inc. (WMT), 5
Weighted standard deviation (WSD) estimator, 23
Wholesale price index (WPI-core), 162
Wholesale price inflation (WPI), 176

X

X-efficiency theory, 269