High-dimensional time series analysis

Valorization

The central theme underlying this thesis is the analysis of high-dimensional time series datasets. The current era is characterized by wide availability of larger and less structured datasets and the process of information extraction from this kind of data demands a drastically different approach that better accommodates these new features. While much progress is being made in the field of high-dimensional statistics in recent years, the analysis of high-dimensional time series in particular merits additional treatment. The appeal of high-dimensional time series analysis stems from the idea of drawing strength from both the time dimension and the (potentially large) cross-sectional dimension to improve model estimates and corresponding forecasts. However, time series analysis in high dimensions comes with its own unique set of challenges. First, the accelerated growth in time series datasets is experienced mostly along the cross-sectional dimension, as changes in information management allow us to extract data from more individuals or agents, but the passing of time puts a strict limit on the growth in the time dimension. Second, even on traditional, smaller datasets the peculiarities of time series such as serial dependencies, non-stationarity and structural breaks call for specialized treatment. The addition of high-dimensionality exacerbates the complexity of the analysis of time series, and the research presented in this thesis aims to contribute to this problem in several ways.

A strong emphasis in this thesis is placed on rigorous and, especially, honest comparison between traditional and state-of-the-art statistical models that have a strong founding in econometric theory. As often the case in transitional periods, it is easy to become convinced by ill-founded claims or idiosyncratic success stories of new methods in exotic applications. Accordingly, the second chapter consolidates seminal and recent literature on prospective statistical methods that are well-suited for forecasting based on high-dimensional time series datasets, and contains elaborate

comparisons of their forecast performance in both controlled and real-life settings. The results provide detailed insights into the relationship between the considered estimators' forecast performance and characteristics of the data (generating process). These insights can serve as a guideline for practitioners facing a forecasting exercise or provide useful benchmarks for the development of new estimators. Furthermore, the value in this chapter is strengthened by our focus on general and realistic data characteristics that are not necessarily specific to a particular field, thus allowing for broad applicability. On a personal level, I particularly hope the research presented in this chapter can be of use to the field of climate science, where temperature forecasts are often based on large datasets of atmospheric measurements containing time series that are characterized by strong (seasonal) dependence over time and cross-sectional dependence due to the proximity between measuring stations. Liberally conjecturing on potential applications, I consider (1) the use of penalized regression to filter out irrelevant atmospheric particles types from the data, (2) using principal component based algorithms to impute missing or faulty measurements and (3) modelling large cointegrated systems of, for example, land and sea temperatures combined with greenhouse gasses as interesting avenues of research that the results in this thesis may be able to contribute towards.

In the third and fourth chapter we develop the Single-equation Penalized Error Correction Selector (SPECS), a novel estimator that combines the traditional approach of cointegration modelling in conditional systems with the dimensionality reduction properties of penalized regression. Ever since its development, cointegration modelling has been an essential tool in the study of economic relationships, with classical examples including purchasing power parity (Juselius and MacDonald, 2004), money demand (Johansen, 1992b) and rational expectation models (Johansen and Swensen, 2004), as well as the study of financial theory such as the present value model of stock prices (Campbell and Shiller, 1987), market efficiency (Dwyer Jr and Wallace, 1992) and numerous market linkages such as that between local gasoline prices and global oil prices (Hendry and Juselius, 2001). More modern applications examine these kind of phenomena on a global scale based on cointegrated panel data (e.g. Westerlund, 2007), where the large cross-sectional dimension calls for specialized high-dimensional methods. If in this cases, the modelling exercise is focussed around only a few variables of interest, SPECS can serve as an automated tool to fit sparse linear single-equation models that incorporate both the long-run and short-run dynamics in the data. As demonstrated in the empirical application of Chaper 3, in which we nowcast Dutch unemployment based on Google Trends data, SPECS is particularly well-suited for the purpose of nowcasting economic variables on such

large macro-economic datasets. Buono et al. (2017) provide an interesting survey of recent studies incorporating various novel sources of big data, such as Google Trends, credit card, social media and stock exchange data, to nowcast macro-economic variables. With the rise of this many sources of big data, the nowcasting potential of SPECS has clearly not been exhausted in the single empirical application considered in Chapter 3.

SPECS may also be used for so-called "Artifical Counterfactual Analysis" (ArCo) in the spirit of Masini and Medeiros (2019). Counterfactual analysis is the examination of treatment effects in the absence of obvious control groups. For example, the highest income tax in the Netherlands was changed in 2001 from 60% to 51%. A natural question to ask is how this has impacted the GDP of the Netherlands. To disentangle the effect of the tax law change and other variables affecting the DGP, one may consider the use of neighbouring economies that were not subjected to this policy change as artificial control groups. Creating artificial controls based on multiple countries and multiple economic indicators quickly gives rise to high-dimensional models, for which SPECS can be considered as a useful estimator. While some theoretical details ought to be worked out, the estimation consistency of SPECS derived in the high-dimensional framework of Chapter 4 is a valuable pre-requisite for ArCo based on SPECS to be considered valid.

For many statical models that form the basis for the determination of economic policy, the ability to perform honest, i.e. uniformly valid, post-selection inference on large (co)integrated datasets is essential. I acknowledge that the thesis does not contribute to this important topic directly. However, from the post-selection inferential tools developed in the stationary world, such as the desparsified lasso (Van de Geer et al., 2014) or post-double selection method (Belloni et al., 2014), it is clear that the theoretical results derived in Chapters 3 and 4 may serve as starting points for the development of novel inferential techniques.

It is worth mentioning that, from the start of the development of SPECS, key considerations have been the intuitiveness of the model and ease of implementation. I believe that the value of an estimator is ultimately derived from its practical usability and the adoption rate among practitioners, no matter how mathematically interesting the underlying theory may be. Not only do I believe that the resulting single-equation model is understandable for a wide audience including non-experts, it is implementable with readily available, off-the-shelf tools including self-written code that I have made publicly available online. Moreover, the relatively low requirements in terms of data pre-processing further reduces the burden on the applied researcher.

In light of the results in Chapter 5 that demonstrate the complexity of controlling (family-wise) error rates of unit root tests in high-dimensions, I consider this automation of the model building process particularly valuable.

Finally, I would like to take the liberty of including an element that is not traditionally part of the valorisation of a thesis: teaching. The accumulation of knowledge throughout a PhD would be worthless to society without its subsequent dissemination. The publication of scientific results tends to reach a rather select audience, whereas knowledge transfer through direct interaction with students often has much farther reaching consequences. Having been lost on my academic path for a while myself, I understand the value of guiding young students in their search for knowledge and self-development. I have had the fortune to teach students from all over the world, with equally varying backgrounds, and made it my goal to connect with them and to convince them of the value of quantitative analysis. Of course, teaching an already excited econometrics student about the power and generality of maximum likelihood estimation has been a great pleasure, but witnessing social science students discovering the value of statistical inference within their fields of interest and, often to their own surprise, becoming excited about statistical theory, felt equally rewarding. I hope I have achieved my goals of inspiring the new generation to pursue their academic interests and I look forward to what the future may bring.