Multivariate state space methods for official statistics and climate modelling

Citation for published version (APA):

Schiavoni, C. (2021). *Multivariate state space methods for official statistics and climate modelling*. [Doctoral Thesis, Maastricht University]. Global Academic Press. https://doi.org/10.26481/dis.20211104cs

Document status and date:
Published: 01/01/2021

DOI:
10.26481/dis.20211104cs

Document Version:
Publisher's PDF, also known as Version of record

Doctoral thesis

MULTIVARIATE STATE SPACE METHODS FOR OFFICIAL STATISTICS AND CLIMATE MODELLING

Caterina Schiavoni

2021

This book was typeset by the author using LaTeX.

Published by Global Academic Press
Cover: Cecilia Matteoli and Luca Baseggio, 2021
ISBN: 978-94-6423-468-8
Printed in The Netherlands by ProefschriftMaken

MULTIVARIATE STATE SPACE METHODS FOR OFFICIAL STATISTICS AND CLIMATE MODELLING

Dissertation

To obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof. Dr. R.M. Letschert,
in accordance with the decision of the Board of Deans,
to be defended in public
on Thursday 4th of November 2021, at 10.00 hours

by

Caterina Schiavoni

Supervisors

Prof. Dr. J.A. van den Brakel
Prof. Dr. F.C. Palm

Co-supervisor

Dr. S.J.M. Smeekes

Assessment Committee

Prof. Dr. A.W. Hecq (Chair)
Dr. N. Baştürk
Prof. Dr. T. Proietti (Tor Vergata University of Rome)
Prof. P.A. Smith (University of Southampton)

This research was financially supported by Statistics Netherlands - Het Centraal Bureau voor de Statistiek (CBS).

Acknowledgments

I would like to thank my supervisors Jan van den Brakel, Stephan Smeekes and Franz Palm for your professionalism, cordiality, availability and patience with me during these years, and Siem Jan Koopman for sharing all your knowledge with a motivating and contagious enthusiasm.

I am grateful to Alain Hecq, Nalan Baştürk, Tommaso Proietti and Paul Smith for the careful reading of this thesis and for having provided detailed feedbacks that helped me improving it.

Karin, Yolanda and Vera, many thanks for all your help with the bureaucracy. Denis, being your teaching assistant has been a truly pleasant and formative experience. I have learnt a lot from you about how to teach econometrics, and I have appreciated your commitment to improve my english (pun intended).

Li, your positive vibe and bald remarks have made sharing the office with you delightful and fun. Benoit, thank you for all the events that you organized but especially for having let me win two tennis sets in three years.

I thank my colleagues at Maastricht University and Statistics Netherlands for the pleasant moments shared at the office, and at the Center for Spatial Data Science of the University of Chicago for having virtually hosted me during the pandemic. I am also very grateful to my friends of Maastricht for having done such a great job in making me escape this thesis every once in a while.

To my friends in Italy and in the rest of Europe: meeting you is always refreshing; growing up with you is a gift. Special thanks go to my family for your support, for putting up with me, and for having brought me pizza at the airport every time I visited home. Hanno, you are brilliant and humble. Your positive and gentle approach to life has been a great source of strength to me. Thank you for your encouragements and cheerfulness.

Caterina Schiavoni
Milan, September 2021

To my four parents and my sister

“I sometimes suspect that inside every data scientist is a kid trying to figure out why his childhood dreams didn’t come true.” —Seth Stephens-Davidowitz, *Everybody Lies*

I wanted to become a paleontologist.

Contents

1
Introduction

State space models are a category of econometric models designed to analyse time series data. The term “state space” originates from engineering and is therefore not helpful, for economists and econometricians, to intuitively understand what these models can be used for. This chapter will provide such understanding.

Compared to (many) other time series econometric models, state space models are particularly suited to estimate latent variables and/or parameters that vary over time. Latent variables can play the role of the seasonality and time trends that are often *latently* present in *observed* time series data, and are generally treated as time-varying in state space models. Parameters of interest can be means, variances or the coefficients that establish relationships among time series. Like latent variables, parameters are also unobserved and, as such, need to be estimated from the data. By allowing parameters to be time-varying, we can model sudden events such as periods of economic distress, legislative changes regarding data collection, or the implementation of new economic/environmental policies.

Latent variables and time-changing parameters constitute the so-called state variables. In order to model state variables as time-varying, it is necessary to pre-specify what type of evolution they follow over time: do they behave wildly or do they fluctuate calmly around a certain mean? State space models are therefore composed of two equations: the observation/measurement equation, which defines the relationship between the observed data and the state variables; the transition/state equation, which determines the (time) dynamics that are followed by the state variables. Both equations are accompanied by their own error terms.

Once the state space model is so set up, the state variables are (computationally) efficiently estimated with a technique called the "Kalman filter", after electrical engineer Rudolf Emil Kálmán. The Kalman filter is then used in order to evaluate the log-likelihood, whose maximisation yields the maximum likelihood estimates of all other parameters that are treated as time-constant[1], i.e., the values of such parameters for which the observed data would have most likely occurred. This efficient estimation of state space models, based on a combination of Kalman filtering and maximum likelihood, is feasible provided the state space model is Gaussian and linear. That is, if the error terms are assumed to be Gaussian/Normally-distributed, and if time-varying (co)variances and other parameters that if treated as state variables can induce nonlinearities in the state space model, are actually not part of the state variables. In case of nonlinear or non-Gaussian state space models, other techniques, than Kalman filtering (and sometimes maximum likelihood), need to be used.

Chapters 2-5 of this thesis propose either a new type of state space model or a new type of econometric technique that deals with a specific case of nonlinear state space models. All of these chapters are accompanied by Monte Carlo simulation studies, which illustrate the finite-sample performance of existing econometric methods in estimating the proposed new models, or of the proposed new methods in estimating existing models. The thesis therefore brings a methodological contribution to the econometric literature. However, the motivation behind the novel econometric models and techniques is always empirical, and the application on real data plays a paramount role in all of

[1] It has to be noted that the Kalman filter can also estimate state variables as static if their transition equations imply a time-constant evolution.

these chapters. So let us now explore the chapters further, starting right from their empirical motivation.

National statistical offices are becoming ever more interested in using different data sources in order to provide more timely and accurate official statistics. The Dutch level of unemployment is an example of an official statistics and it plays a central role in this thesis. Statistics Netherlands, in particular, is already making use of state space models in order to estimate the Dutch unemployment, by treating the latter as a latent variable (or time-varying mean) of survey-based time series data. However, surveys do not represent the only data source that is informative about unemployment. Claimant counts, which are the number of people claiming unemployment benefits, and are registered, in the Netherlands, by the Ministry of Social Affairs and Employment, also constitute a good time series indicator about unemployment. Combining both data sources can already result in more accurate estimates of unemployment. Although surveys have the advantage of providing accurate information about unemployment, they are a rather expensive and slow way of collecting data. The latter applies also to claimant counts. Indeed, both types of data are collected on a monthly basis but are published with a delay of one month. This means we have to wait until next month before we can get a good estimate for the current monthly level of unemployment, based on these two data sources; we can only try to predict its current value (in other words, "nowcast" it). And here is where the famous Big Data come into play, since they are generally available in real time and at high frequency, and can so be useful in achieving more timely estimates of variables of interest. Contrary to the types of data discussed so far, Big Data are a rather unstructured type of data source: they are "Big" and therefore rich in information, but they also contain lots of noise that needs to be filtered out. Google Trends, which are time series of Google searches, are an example of Big Data. They can be informative of the unemployment if one thinks of searches that can be made by unemployed people, as "job search" or "I am looking for a job". Clearly, there are many such Google search queries and the resulting data set could be categorized as "Big" indeed. However, there are some terms, like "cv", which in the Dutch stands for "curriculum vitae" but also "central heating", and which can after all be unrelated to the Dutch unemployment; hence the noise affecting the same data. In

Chapter 2 we extend the above-mentioned state space model used at Statistics Netherlands, by employing the three types of data sources just discussed, and we investigate to which extent we can so improve the estimation and nowcast accuracy of the Dutch unemployment, over only using survey-based data. The noise of Google Trends is filtered out by means of high-dimensional (i.e., designed to deal with Big Data) econometric techniques: penalised regressions and factor models. The methodological way to integrate and exploit high-dimensional times series in state space models is new and preserves the linearity of the model. Hence, we inspect how the Kalman filter performs in nowcasting the state variables of our proposed high-dimensional state space model.

Combining several time series from different data sources into a single model requires linking them by means of parameters. We do so via correlation parameters that enter the covariance matrix of the transition equation's error terms. Correlations are large in magnitude if the series are strongly related and hence all informative about the unemployment. In the framework of Chapter 2 we implicitly assumed that these correlation parameters are time-constant, meaning that the relationships among time series are not changing over time. Nonetheless, this assumption is rather strong and unlikely to hold in practice, especially if the considered time span is long. For instance, the economic crisis of 2008 is often found, in the econometric literature, to trigger variations in time series' parameters. Likewise, data collection can be subject to legislative changes that can induce modifications in the relations among time series that aim at measuring the same variable, but are based on different data sources. In Chapter 3 we therefore let the correlation parameter linking survey-based and claimant count data vary over time. We investigate two methods to model such time changing relationship. The first one is based on cubic splines, which is an interpolation method, and which eases the estimation of the state space model by retaining the linearity of the latter. However, the method requires the unrealistic assumption that the correlation is varying smoothly over time. The second technique treats the time-varying correlation as an additional state variable, which therefore has its own source of error and triggers the nonlinearity of the state space model. We have already learnt that in this case we cannot use Kalman filtering and maximum likelihood to estimate the state space model. We therefore propose an estimation method that replaces the latter by particle filtering and indirect inference, respectively. The particle filter, as its name suggests, is another filtering method to estimate the state variables. The indirect inference approach makes use of an easy-to-estimate linear state space model, which approximates the nonlinear one, in order to estimate the static parameters of the latter. In our case the model that makes use of cubic splines is the approximate model, and this represents a novelty in the literature about indirect inference. Our proposed methodology is more complex and computationally expensive than the one purely based on cubic splines, but it requires much less restrictive assumptions on the evolution of the correlation over time, and subsequent more realistic estimates of the latter. The chapter therefore contributes to the methodological literature on the estimation of nonlinear state space models. We compare the performance of the two methods in estimating the time-varying correlation. We also investigate whether modelling the correlation as time-varying, instead of time constant, allows us to achieve even more accurate and realistic estimates of Dutch unemployment.

But why did we not employ already existing econometric methods designed to estimate nonlinear state space models, such as importance sampling or the Extended Kalman filter? Because we did not manage to make them work. The type of nonlinearity we dealt with in Chapter 3, is particularly nasty since it is present in the covariance matrix of the transition equation's error terms. Also the Extended Kalman filter and importance sampling use approximate linear state space models in order to estimate the nonlinear one, but they are more suited to deal with nonlinearities that are either located in the observation equation of the state space model, or that do not involve the error terms of the transition equation. We therefore did try to shift the nonlinear part of our model to a location that would facilitate the implementation of the two above-mentioned existing methods, but without any success. These attempts are anyway documented in Chapter 4. Moreover, cubic splines do not represent the only existing method to model parameters as time-varying, while preserving the linearity of the state space models. The score-driven approach can alternatively be employed to do so. The idea behind this method is to let the time-varying correlation depend only on its past values and past values of the score of the log-likelihood. The latter measures the sensitivity of the likelihood with respect to changes in the correlation, and is therefore informative to predict future values of the correlation parameter. In Chapter 4 we also show how to implement the score-driven technique, in order to estimate the time-varying correlation. Although the implementation of this approach is feasible, the results are not satisfactory: the method is indeed barely capable of detecting time-changes in the correlation parameter and therefore does not represent a good candidate to model time-varying state correlations in state space models. Hence, rather than a publishable paper, Chapter 4 collects some notes that have been taken by the author while exploring unsuccessful approaches to answer the research questions of Chapter 3. As such, it is an illustration of the research process in taking several dead ends before ending up in the right track.

Chapters 2-4 are all focused on shaping new econometric models and methods that can be mainly (but not only) employed for the production of more accurate (and timely) official statistics. Chapter 5 instead takes off from this subject and lands on another topical and growing field: environmental studies. Climate change is indeed increasingly catching the attention of some econometricians, who are employing econometric techniques in order to understand and model it, and eventually contribute to the implementation and evaluation of policies aimed at combating it. Along these lines, in Chapter 5 we use a state space approach in order to model and forecast time series of Dutch regional concentrations of nitrogen dioxide (NO_2), which is an air pollutant. It is primarily emitted during anthropogenic combustion processes, of which on road vehicles represent the main source. NO_2 is responsible for the creation of secondary aerosols and pollutants and hence indirectly affects the climate. Additionally, it is directly detrimental for ecosystems and human health, making it an important pollutant to analyse. Environmental variables are generally characterised by spatial dependence. That is, regional air pollution can depend on and affect the pollution of neighbouring regions, because of its transportation by the wind. Our proposed state space model has therefore the novel feature of taking such spatial interactions into account. In building this model, we also control for meteorological factors that may affect the concentrations of the pollutant under study. Moreover, we let the effect of traffic intensity on such concentrations vary over time, in order to capture potential changes on the efficiency of vehicles that might in turn reduce the contribution of the latter to the quantity of air pollution, over time. The state space structure of the model further allows us to innovatively capture time-varying border effects, which represent differences between peripheral and inland regions, via state variables. Finally, we use this model to forecast NO_2 concentrations for different scenarios of traffic intensity, showing its potential in evaluating pollution-reduction policies. Our proposed spatial state space model is linear and Gaussian, hence we evaluate the performance of the usual Kalman filter and maximum likelihood methods in estimating its state variables (which include time-varying parameters) and static parameters, respectively.

Since state space models are designed to fit time series data, the number of time observations plays a crucial role. Throughout the thesis, we refer to this number as the "sample size". It has to be large enough in order to allow us to estimate state space models (and establish the properties of the estimation methods employed), and of course the larger it is, the better we can do so.

Recall that in Chapters 2-4 we make use of different data sources to estimate the same latent variable of interest. Moreover, in Chapter 5 we add a cross-sectional dimension to the model, represented by the Dutch regions, which entails the joint modelling of the time series of NO_2 concentrations for every region. Therefore, all state space models employed in this thesis are *multivariate*, as they fit several time series at the same time. However, the number of these series does not have to be large enough to guarantee a feasible estimation of the model, contrary to the "sample size". This paragraph should provide the last piece of the puzzle in order to understand the thesis' title, and fully prepare us to dive into the chapters in more detail.

2
A dynamic factor model approach to incorporate Big Data in state space models for official statistics

Adapted from: Schiavoni et al. (2021b).

Abstract

In this chapter we consider estimation of unobserved components in state space models using a dynamic factor approach to incorporate auxiliary information from high-dimensional data sources. We apply the methodology to unemployment estimation as done by Statistics Netherlands, who use a multivariate state space model to produce monthly figures for unemployment using series observed with the labour force survey (LFS). We extend the model by including auxiliary series of Google Trends about job-search and economic uncertainty, and claimant counts, partially observed at higher frequencies. Our factor model allows for nowcasting the variable of interest, providing reliable unemployment estimates in real-time before LFS data become available.

2.1 Introduction

There is an increasing interest among national statistical institutes (NSIs) to use data that are generated as a by-product of processes not directly related to statistical production purposes in the production of official statistics. Such data sources are sometimes referred to as “Big Data”; examples are time and location of network activity available from mobile phone companies, social media messages from Twitter and Facebook, sensor data, and internet search behaviour from Google Trends. A common problem with this type of data sources is that they are likely selective with respect to an intended target population. If such data sources are directly used to produce statistical information, then the potential selection bias of these data sources must be accounted for, which is often a hard task since Big Data sources are often noisy and (although they contain lots of information) they generally contain no auxiliary variables, which are required for bias correction. These problems can be circumvented by using them as covariates in model-based inference procedures to make precise detailed and timely survey estimates, since they come at a high frequency and are therefore very timely. These techniques are known in the literature as small area estimation and nowcasting (Rao & Molina, 2015).

Official statistics are generally based on repeated samples. Therefore multivariate time series models are potentially fruitful to improve the precision and timeliness of domain estimates with survey data obtained in preceding reference periods and other domains. The predictive power of these models can be further improved by incorporating auxiliary series that are related with the target series observed with a repeated survey.

In this chapter we investigate how auxiliary series derived from big data sources and registers can be combined with time series observed with repeated samples in high dimensional multivariate structural time series (STS) models. We consider Google Trends and claimant counts as auxiliary series for monthly unemployment estimates observed with a continuously conducted sample survey. Big Data sources have the problem that they are noisy and potentially (partly) irrelevant, and, as such, care must be taken when using them for the production of official statistics. We show that, by using a dynamic factor model in state space form, relevant information can be extracted from such auxiliary high-dimensional data sources, while guarding against the inclusion of irrelevant data.

Statistical information about a country's labour force is generally obtained from labour force surveys, since the required information is not available from registrations or other administrative data sources. The Dutch labour force survey (LFS) is based on a rotating panel design, where monthly household samples are observed five times at quarterly intervals. These figures are, however, considered too volatile to produce sufficiently reliable monthly estimates for the employed and unemployed labour force at monthly frequency. For this reason Statistics Netherlands estimates monthly unemployment figures, together with its change, as unobserved components in a state space model where the observed series come from the monthly Dutch LFS, using a model originally proposed by Pfeffermann (1991). This method improves the precision of the monthly estimates for unemployment with sample information from previous periods, and can therefore be seen as a form of small area estimation. In addition it accounts for rotation group bias (Bailar, 1975), serial correlation due to partial sample overlap, and discontinuities due to several major survey redesigns (van den Brakel & Krieg, 2015).

Time series estimates for the unemployment can be further improved by including related auxiliary series. The purpose is twofold. First, auxiliary series can further improve the precision of the time series predictions. In this regard, Harvey and Chung (2000) propose a bivariate state space model to combine a univariate series of the monthly unemployed labour force derived from the UK LFS, with the univariate auxiliary series of claimant counts. The latter series represents the number of people claiming unemployment benefits. It is an administrative source, which is not available for every country, and, as for the Netherlands, it can be affected by the same publication delay of the labour force series. Second, auxiliary series derived from Big Data sources like Google Trends are generally available at a higher frequency than the monthly series of the LFS. Combining both series in a time series model allows us to make early predictions for the survey outcomes in real-time at the moment that the outcomes for the auxiliary series are available, but the survey data not yet, which is in the literature known as *nowcasting*, in other words, "forecasting the present".

In this chapter, we extend the state space model used by Statistics Netherlands in order to combine the survey data with the claimant counts and the high-dimensional auxiliary series of Google Trends about job-search and economic uncertainty, as they could yield more information than a univariate one, which is not affected by publication lags and that can be observed at a higher frequency than the labour force series.

This chapter contributes to the existing literature by proposing a method to include a high-dimensional auxiliary series in a state space model in order to improve the (real-time) estimation of unobserved components. The model accounts for the rotating panel design underlying the sample survey series, combines series observed at different frequencies, and deals with missing observations at the end of the sample due to publication delays. It handles the curse of dimensionality that arises from including a large number of series related to the unobserved components, by extracting their common factors.

Besides claimant counts, the majority of the information related to unemployment is nowadays available on the internet; from job advertisements to resumé's templates and websites of recruitment agencies. We therefore follow the idea originating in Choi and Varian (2009), Askitas and Zimmermann (2009) and Suhoy (2009) of using terms related to job and economic uncertainty, searched on Google in the Netherlands. Since 2004, these time series are freely downloadable in real-time from the Google Trends tool, at a monthly or higher frequency. As from the onset it is unclear which search terms are relevant, and if so, to what extent, care must be taken not to model spurious relationships with regards to the labour force series of interest, which could have a detrimental effect on the estimation of unemployment, such as happened for the widely publicized case of Google Flu Trends (Lazer et al., 2014).

Our method allows us to exploit the high-frequency and/or real-time information of the auxiliary series, and to use it in order to nowcast the unemployment, before the publication of labour force estimates. As the number of search terms related to unemployment can easily become large, we employ the two-step estimator of Doz et al. (2011), which combines factor models with the Kalman filter, to deal both with the high-dimensionality of the auxiliary series, and with the estimation of the state space model. The above-mentioned estimator is generally used to improve the nowcast of variables that are treated as observed, such as GDP (see Giannone et al. (2008) and Hindrayanto et al. (2016) for applications to the US and the euro area). D'Amuri and Marcucci (2017), Naccarato et al. (2018), and Maas (2019) are all recent studies that use Google Trends to nowcast and forecast the unemployment, by treating it as a *known* dependent variable in time series models where the Google searches are part of the explanatory variables. Nonetheless, unemployment is not actually observed, and to the best of our knowledge, this chapter is the first one to use Google Trends in order to nowcast the unemployment in a model setting that treats it variable as *unobserved*.

We evaluate the performance of our proposed method via Monte Carlo simulations and find that our method can yield large improvements in terms of Mean Squared Forecast Error (MSFE) of the unobserved components' nowcasts. We then assess whether the accuracy of the unemployment's estimation and nowcast improves with our high-dimensional state space model, respectively from in-sample and out-of-sample results. The latter consists of a recursive nowcast. We do not venture into forecasting exercises as Google Trends are considered to be more helpful in predicting the present rather than the future of economic activities (Choi & Varian, 2012). We conclude that Google Trends can significantly improve the fit of the model, although the magnitude of these improvements is sensitive to aspects of the data and the model specification, such as the frequency of observation of the Google Trends, the number of Google Trends' factors included in the model, and the level of estimation accuracy provided by the first step of the two-step estimation procedure.

The remainder of the chapter is organized as follows. Section 2.2 discusses the data used in the empirical analysis. Section 2.3.1 describes the state space model that is currently used by Statistics Netherlands to estimate unemployment. Section 2.3.2 focuses on our proposed method to include a high-dimensional auxiliary series in the aforementioned model. Sections 2.4 and 2.5 report, respectively, the simulation and empirical results for our method. Section 2.6 concludes.

2.2 Data

The Dutch LFS is conducted as follows. Each month a stratified two-stage cluster design of addresses is selected. Strata are formed by geographical regions. Municipalities are considered as primary sampling units and addresses as secondary sampling units. All households residing at an address are included in the sample with a maximum of three (in the Netherlands there is generally one household per address). All household members with age of 16 or older are interviewed. Since October 1999, the LFS has been conducted as a rotating panel design. Each month a new sample, drawn according to the above-mentioned design, enters the panel and is interviewed five times at quarterly intervals. The sample that is interviewed for the j^{th} time is called the j^{th} wave of the panel, $j = 1, \ldots, 5$. After the fifth interview, the sample of households leaves the panel. This rotation design implies that in each month five independent samples are observed. The generalized regression (GREG, i.e., design-based) estimator (Särndal et al., 1992) is used to obtain five independent direct estimates for the unemployed labour force, which is defined as a population total. This generates over time a five-dimensional time series of the unemployed labour force. Table 2.1 provides a visualization for the rotation panel design of the Dutch LFS.

quarter

month

A B C	D E F	G H I	J K L	M N O	P Q R	} wave 1
	A B C	D E F	G H I	J K L	M N O	} wave 2
		A B C	D E F	G H I	J K L	} wave 3
			A B C	D E F	G H I	} wave 4
				A B C	D E F	} wave 5

Table 2.1: Visualization for the rotation panel design of the Dutch LFS. Each capital letter represents a sample. Every month a new sample enters the panel and is interviewed five times at a quarterly frequency. After the fifth interview, the sample of households leaves the panel.

Rotating panel designs generally suffer from Rotation Group Bias (RGB), which refers to the phenomenon that there are systematic differences among the observations in the subsequent waves (Bailar, 1975). In the Dutch LFS the estimates for the unemployment based on the first wave are indeed systematically larger compared to the estimates based on the follow-up waves (van den Brakel & Krieg, 2015). This is the net results of different factors:

- Selective nonresponse among the subsequent waves, i.e., panel attrition.
- Systematic differences due to different data collection models that are applied to the waves. Until 2010 data collection in the first wave was based on face-to-face interviewing. Between 2010 and 2012 data collection in the first wave was based on telephone interviewing for households for which a telephone number of a landline telephone connection was available and face-to-face interviewing for the remaining households. After 2012 data collection in the first wave was based on a sequential mixed mode design that starts with Web interviewing with a follow up using telephone interviewing and face-to-face interviewing. Data collection in the follow-up waves is based on telephone interviewing only.
- Differences in wording and questionnaire design used in the waves. In the first wave a block of questions is used to verify the status of the respondent in the labour force. In the follow-up waves the questionnaire focuses on differences that occurred compared to the previous interview, instead of repeating the battery of questions.
- Panel conditioning effects, i.e., systematic changes in the behaviour of the respondents. For example, questions about activities to find a job in the first wave might increase the search activities of the unemployed respondents in the panel. Respondents might also systematically adjust their answers in the follow-up waves, since they learn how to keep the routing through the questionnaire as short as possible.

The Dutch labour force is subject to a one-month publication delay, which means that the sample estimates for month t become available in month $t+1$. In order to have more timely and precise estimates of unemployment, we extend the model by including, respectively, auxiliary series of weekly/monthly Google Trends about job-search and economic uncertainty, and monthly claimant counts, in the Netherlands.

Claimant counts are the number of registered people that receive unemployment benefits. The claimant counts for month t become available in month $t + 1$.

Google Trends are indexes of search activity. Each index measures the fraction of queries that include the term in question in the chosen geography at a particular time, relative to the total number of queries at that time. The maximum value of the index is set to be 100. According to the length of the selected period, the data can be downloaded at either monthly, weekly, or higher frequencies. The series are standardized according to the chosen period and their values can therefore vary according to the period's length (Stephens-Davidowitz & Varian, 2015). We use weekly and monthly Google Trends for each search term. Google Trends are available in real-time (i.e., they are available in period t for period t, independently on whether the period is a week or a month).

The list of Google search terms used in the empirical analysis of this chapter, together with their translation/explanation, is reported in Tables 2.B.1 and 2.B.2. A first set of terms (which is the one used in a previous version of this chapter) was chosen by thinking of queries that could be made by unemployed people in the Netherlands. The rest of the terms have been chosen by using the Google Correlate tool and selecting the queries that are highly correlated to each term of the initial set, and that have a meaningful relation to unemployment and, more generally, economic uncertainty[1].

Figure 2.1 displays the time series of the five waves of the unemployed labour force, together with the claimant counts and an example of a job-related Google query. They show a similar behaviour over time, which already shows the potential of using this auxiliary information in estimating unemployment.

[1] Later in the chapter we mention that we need non-stationary (e.g., persistent) Google Trends for our model. Correlations between non-stationary series can be spurious, and in this respect Google Correlate is not an ideal tool in order to choose search terms. In section 2.5 we explain how to circumvent this problem.

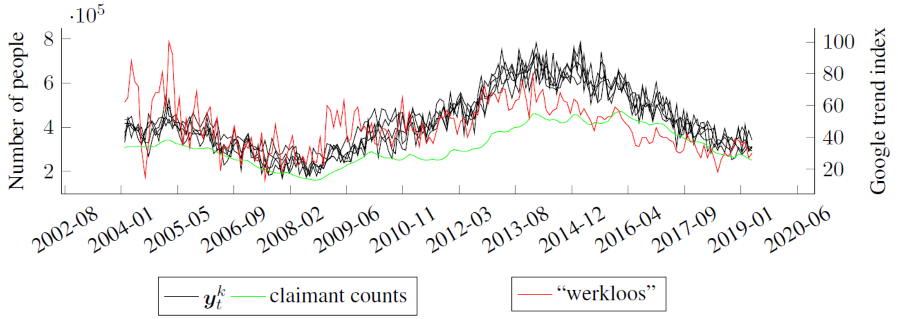


Figure 2.1: Monthly time series of the five waves of the Dutch unemployed labour force ($\boldsymbol{y}_t^k$), the claimant counts, and the Google search term "werkloos", which means "unemployed", in the Netherlands. The period starts in January 2004 and ends in May 2019.

2.3 The Dutch labour force model and extensions

We first describe the model in use at Statistics Netherlands in Section 2.3.1. Next we explain how high-dimensional auxiliary series can be added to this model in Section 2.3.2.

2.3.1 The Dutch labour force model

The monthly sample size[2] of the Dutch LFS is too small to produce sufficiently precise estimates directly. In the past, rolling quarterly figures were published at a monthly frequency. This has the obvious drawback that published figures are unnecessarily delayed since the reference period is the mid month of the rolling quarter. Also, real monthly seasonal effects are smoothed over the rolling quarter. Another problem that arose after the change from

[2] This sample size corresponds to the number of people that have been surveyed in a given month, and that is used for a first estimation of Dutch monthly unemployment. However, it is different from T which, as mentioned in the Introduction, is the sample size that we need to be large enough in order to accurately re-estimate Dutch unemployment with a state space model.

a cross-sectional survey to a rotating panel design in 2000, was that the effects of RGB became visible in the labour force figures. Both problems are solved with a structural time series (STS) model, that has been used by Statistics Netherlands, since 2010, for the production of monthly statistics about the Dutch labour force (van den Brakel & Krieg, 2015). In a STS model, an observed series is decomposed into several unobserved components, such as a trend, a seasonal component, one or more cycles with a period longer than one year, regression components, and a white noise component. After writing an STS model in the state space form, the Kalman filter can be applied in order to estimate the unobserved components. See Durbin and Koopman (2012) for an introduction to STS modelling.

Let $y^k_{j,t}$ denote the GREG estimate for the unemployment in month t based on the sample observed in wave j. Now $\boldsymbol{y}^k_t = (y^k_{1,t}, \dots, y^k_{5,t})$ denotes the vector with the five GREG estimates for unemployment in month t. The $y^k_{j,t}$ are treated as five distinct time series in a five dimensional time series model in order to account for the rotation group bias. The superscript $k > 1$ indicates that the vector is observed at the low frequency. We need this notation (see e.g. Bańbura et al., 2013) to distinguish between series observed at different frequencies, because later on we will make use of Google Trends which are available on a weekly basis. If $\boldsymbol{y}^k_t$ is observed at the monthly frequency, as in the case of the unemployed labour force, then $k = 4, 5$ if the high frequency series is observed at the weekly frequency, since a month can have either 4 or 5 weeks.

The unemployment is estimated, with the Kalman filter, as a state variable in a state space model where $\boldsymbol{y}^k_t$ represents the observed series. The measurement equation takes the form (Pfeffermann, 1991; van den Brakel & Krieg, 2009):

$$\boldsymbol{y}^k_t = \boldsymbol{\imath}_5 \theta^{k,y}_t + \boldsymbol{\lambda}^k_t + \boldsymbol{e}^k_t. \tag{2.3.1}$$

where $\boldsymbol{\imath}_5$ is a 5-dimensional vector of ones, and $\theta^{k,y}_t$, i.e. the unemployment, is the common population parameter among the five-dimensional waves of the unemployed labour force. It is composed of the level of a trend (L_t) and a seasonal component (S_t):

$$\theta_t^{k,y} = L_t^{k,y} + S_t^{k,y}.$$

The transition equations for the level (L_t) and the slope (R_t) of the trend are, respectively:

$$\begin{aligned} L_t^{k,y} &= L_{t-1}^{k,y} + R_{t-1}^{k,y}, \\ R_t^{k,y} &= R_{t-1}^{k,y} + \eta_{R,t}^{k,y}, \quad \eta_{R,t}^{k,y} \sim N\left(0, \sigma_{R,y}^2\right), \end{aligned}$$

which characterize a smooth trend model. This implies that the level of the trend is integrated of order 2, denoted as $I(2)$, which means that the series of the level is stationary (i.e., mean-reverting) after taking two times successive differences. The slope of the trend, $R_t^{k,y}$, is a first-order integrated series, denoted as $I(1)$. This state variable represents the change in the level of the trend, $L_t^{k,y}$, and not in the unemployment, $\theta_t^{k,y}$, directly. Nevertheless, since the $I(2)$ property of the unemployment is driven by its trend, and not by its seasonal component, the change in $\theta_t^{k,y}$ will also mainly be captured by $R_t^{k,y}$, and we can therefore consider the latter as a proxy for the change in unemployment. A previous version of the model contained an innovation term for the population parameter $\theta_t^{k,y}$. However, the maximum likelihood estimate for its variance tended to be zero and Bollineni-Balabay et al. (2017) showed via simulations that it is better to not include this term in the model.

The trigonometric stochastic seasonal component allows for the seasonality to vary over time, and it is modeled as in Durbin and Koopman (2012, Chapter 3):

$$\begin{aligned} S_t^{k,y} &= \sum_{l=1}^{6} S_{l,t}^{k,y}, \\ \begin{pmatrix} S_{l,t}^{k,y} \\ S_{l,t}^{*k,y} \end{pmatrix} &= \begin{bmatrix} \cos(h_l) & \sin(h_l) \\ -\sin(h_l) & \cos(h_l) \end{bmatrix} \begin{pmatrix} S_{l,t-1}^{k,y} \\ S_{l,t-1}^{*k,y} \end{pmatrix} + \begin{pmatrix} \eta_{\omega,l,t}^{k,y} \\ \eta_{\omega,l,t}^{*k,y} \end{pmatrix}, \\ \begin{pmatrix} \eta_{\omega,l,t}^{k,y} \\ \eta_{\omega,l,t}^{*k,y} \end{pmatrix} &\sim N\left(\mathbf{0}, \sigma_{\omega,y}^2 \boldsymbol{I}_2\right), \end{aligned}$$

where $h_l = \frac{\pi l}{6}$, for $l = 1, \ldots, 6$, and $\boldsymbol{I}_2$ is a 2×2 identity matrix.

The second component in equation (2.3.1), $\boldsymbol{\lambda}_t^k = (\lambda_{1,t}^k, \ldots, \lambda_{5,t}^k)^t$, accounts for the RGB. Based on the factors that contribute to the RGB, as mentioned in Section 2.2, the response observed in the first wave is assumed to be the most reliable one and not to be affected by the RGB (van den Brakel & Krieg, 2009). Therefore it is assumed that $\lambda_{1,t}^k = 0$. The remaining four components in $\boldsymbol{\lambda}_t^k$ are random walks that capture time-dependent differences between the follow-up waves with respect to the first wave[3]

$$\begin{aligned}
&\lambda_{1,t}^k = 0, \\
&\lambda_{j,t}^k = \lambda_{j,t-1}^k + \eta_{\lambda,j,t}^k, \quad \eta_{\lambda,j,t}^k \sim N\left(0, \sigma_\lambda^2\right), \quad j = 2, \ldots, 5.
\end{aligned}$$

As a result the Kalman filter estimates for $\theta_t^{k,y}$ in (2.3.1) are benchmarked to the level of the GREG series of the first wave.

The third component in equation (2.3.1), $\boldsymbol{e}_t^k = (e_{1,t}^k, \ldots, e_{5,t}^k)^t$, models the autocorrelation among the survey errors ($e_{j,t}^k$) in the follow-up waves due to the sample overlap of the rotating panel design. In order to account for this autocorrelation, the survey errors are treated as state variables, which follow the transition equation below.

$$\begin{aligned}
&e_{j,t}^k = c_{j,t}\tilde{e}_{j,t}^k, \quad c_{j,t} = \sqrt{\widehat{\text{var}}\left(y_{j,t}^k\right)}, \quad j = 1, \ldots, 5, \\
&\tilde{e}_{1,t}^k \sim N\left(0, \sigma_{\nu_1}^2\right), \\
&\tilde{e}_{j,t}^k = \delta\tilde{e}_{j-1,t-3}^k + \nu_{j,t}^k, \quad \nu_{j,t}^k \sim N\left(0, \sigma_{\nu_j}^2\right), \quad j = 2, \ldots, 5, \quad |\delta| < 1. \\
&\text{var}\left(\tilde{e}_{j,t}^k\right) = \sigma_{\nu_j}^2 / \left(1 - \delta^2\right), \quad j = 2, \ldots, 5,
\end{aligned} \tag{2.3.2}$$

[3] The choice (made by Statistics Netherlands) to model the state variables for the RGB as random walks was arbitrary. In principle it should be possible to first model the contrasts as AR(1) processes and then test whether the autoregressive coefficients are equal to 1, for instance with a likelihood ratio test.

with $\widehat{\text{var}}\left(y_{j,t}^{k}\right)$ being the design variance of the GREG estimates $y_{j,t}^{k}$. The scaled sampling errors, $\tilde{e}_{j,t}^{k}$, for $j = 1, \ldots, 5$, account for the serial autocorrelation induced by the sampling overlap of the rotating panel. Samples in the first wave are observed for the first time and therefore its survey errors are not autocorrelated with survey errors of previous periods. The survey errors of the second to fifth wave are correlated with the survey errors of the previous wave three months before. Based on the approach proposed by Pfeffermann et al. (1998), van den Brakel and Krieg (2009) motivate that these survey errors should be modelled as an AR(3) process, without including the first two lags. Moreover, the survey errors of all waves are assumed to be proportional to the standard error of the GREG estimates. In this way the model accounts for heterogeneity in the variances of the survey errors, which are caused by changing sample sizes over time. As a result the maximum likelihood estimates of the variances of the scaled sampling errors, $\sigma_{\nu_j}^{2}$, will have values approximately equal to one.

The structural time series model (2.3.1) as well as the models proposed in the following sections are fitted with the Kalman filter after putting the model in state space form. We use an exact initialization for the initial values of the state variables of the sampling error, and a diffuse initialization for the other state variables. It is common to call *hyperparameters* the parameters that define the stochastic properties of the measurement equation and the transition equation of the state space model. These are the parameters that are assumed to be known in the Kalman filter (Durbin & Koopman, 2012, Chapter 2). In our case the hyperparameters are δ and all the parameters that enter the covariance matrices of the innovations. These hyperparameters are estimated by maximum likelihood using the Broyden-Fletcher-Goldfarh-Shanno (BFGS) optimization algorithm. The additional uncertainty of using maximum likelihood estimates for the hyperparameters in the Kalman filter is ignored in the standard errors of the filtered state variables. Since the observed time series contains 185 monthly periods, this additional uncertainty can be ignored. See also Bollineni-Balabay et al. (2017) for details. Both the simulation and estimation results in Sections 2.4 and 2.5 are obtained using the statistical software R.

Assuming normality of the innovations is common in state space models because the hyperparameters of the model are estimated by maximizing a Gaussian log-likelihood which is evaluated by the Kalman filter. Moreover, under normality, the Kalman filter yields the minimum variance unbiased estimator of the state variables. Nonetheless, as long as the state space model is linear, if the true distribution of the error terms is non-Gaussian, then the Kalman filter still provides the minimum variance *linear* unbiased estimator of the state variables (Durbin & Koopman, 2012, Chapter 4). In this case we can further rely on quasi maximum likelihood (QML) theory in order to perform inference based on the QML estimates of the hyperparameters. This means that the hyperparameters can still be consistently estimated by maximizing the Gaussian log-likelihood (or in general, as Gourieroux et al. (1984) argue, a density function that belongs to the family of linear exponential distributions), but we shall use, if needed, the appropriate expression for the covariance matrix of the QML estimators, which should capture the additional uncertainty caused by the model's misspecification (Hamilton, 1994, Chapter 13). In Appendix 2.C we conduct a Monte Carlo simulations study and find that deviations from normality are not of concern for the performance our method.

This time series model addresses and solves the mentioned problems with small sample sizes and rotation group bias. Every month a filtered estimate for the trend ($L_t^{k,y}$) and the population parameter, which is defined as the filtered trend plus the filtered seasonal effect ($\theta_t^{k,y} = L_t^{k,y} + S_t^{k,y}$), are published in month $t+1$. The time series model uses sample information from previous months in order to obtain more stable estimates. The estimates account for RGB by benchmarking the estimates for $L_t^{k,y}$ and $\theta_t^{k,y}$ to the level of the first wave, which makes them comparable with the outcomes obtained under the cross-sectional design before 2000.

We now introduce some further notation to distinguish between in-sample estimates and out-of-sample forecasts. In the case of in-sample estimates, $\hat{\theta}_{t|\Omega_t}^{k,y}$ denotes the filtered estimate of the population parameter $\theta_t^{k,y}$, assuming that all data for time t is released and available at time t. We therefore condition on the information set Ω_t which does not contain any missing data at time t. In the case of out-of-sample forecasts, we condition on the data set Ω_t^- that is actually available in real time at time t. For instance, $\boldsymbol{y}_t^k$ only gets published during moth $t+1$, and is therefore not available yet at time t, and not part of Ω_t^-. Thus $\hat{\theta}_{t|\Omega_t^-}^{k,y}$ is the filtered forecast for $\theta_t^{k,y}$, based on the information that is available at time t. Under model (2.3.1), which does not contain auxiliary information other than the labour force series, $\hat{\theta}_{t|\Omega_t^-}^{k,y}$ is in fact the one-step-ahead prediction $\hat{\theta}_{t|\Omega_{t-1}}^{k,y}$, since $\boldsymbol{y}_t^k$ is not available yet in month t, but $\boldsymbol{y}_{t-1}^k$ is; therefore, $\Omega_t^- = \Omega_{t-1} = \{\boldsymbol{y}_{t-1}^k, \boldsymbol{y}_{t-2}^k, \ldots\}$.

2.3.2 Including high-dimensional auxiliary series

To improve precision and timeliness of the monthly unemployment figures, we extend the labour force model by including auxiliary series of weekly/monthly Google Trends about job-search and economic uncertainty, and monthly claimant counts, in the Netherlands. Since the claimant counts for month t become available in month $t+1$, it is anticipated that this auxiliary series is particularly useful to further improve the precision of the trend and population parameter estimates after finalizing the data collection for reference month t. The Google Trends already come at a higher frequency during the reference month t. It is therefore anticipated that these auxiliary series can be used to make first provisional estimates for the trend and the population parameter of the LFS during month t, when the sample estimates $\boldsymbol{y}_t^k$ are not available, but the Google Trends become available on a weekly basis.

Weekly and monthly Google Trends are throughout the chapter denoted by $\boldsymbol{x}_t^{GT}$ and $\boldsymbol{x}_t^{k,GT}$, respectively. We denote the dimension of the vector $\boldsymbol{x}_t^{GT}$ by n, which can be large. In addition, we can expect the Google Trends to be very noisy, such that the signal about unemployment contained in them is weak. We therefore need to address the high-dimensionality of these auxiliary series, in order to make the dimension of our state space model manageable for estimation, and extract the relevant information from these series. For this purpose we employ a factor model which achieves both by retaining the information of these time series in a few common factors.

Moreover, when dealing with mixed frequency variables and with publication delays, we can encounter "jagged edge" datasets, which have missing values at the end of the sample period. The Kalman filter computes a prediction for the unobserved components in the presence of missing observations for the respective observable variables.

The two-step estimator by Doz et al. (2011) combines factor models with the Kalman filter and hence addresses both of these issues. In the remainder of this section we explain how this estimator can be employed to nowcast the lower-frequency unobserved components of the labour force model using information from higher-frequency or real-time auxiliary series.

We consider the following state space representation of the dynamic factor model for the Google Trends data, with respective measurement and transition equations, as we would like to link it to the state space model used to estimate the unemployment (2.3.1):

$$\begin{aligned} \boldsymbol{x}_t^{GT} &= \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Psi}) \\ \boldsymbol{f}_t &= \boldsymbol{f}_{t-1} + \boldsymbol{u}_t, \quad \boldsymbol{u}_t \sim N(\mathbf{0}, \boldsymbol{I}_r), \end{aligned} \tag{2.3.3}$$

where $\boldsymbol{x}_t^{GT}$ is a $n \times 1$ vector of observed series, $\boldsymbol{f}_t$ is a $r \times 1$ vector of latent factors with $r \ll n$, $\boldsymbol{\Lambda}$ is a $n \times r$ matrix of factor loadings, $\boldsymbol{\varepsilon}_t$ is the $n \times 1$ vector of idiosyncratic components and $\boldsymbol{\Psi}$ its $n \times n$ covariance matrix; $\boldsymbol{u}_t$ is the $r \times 1$ vector of factors' innovations and $\boldsymbol{I}_r$ is a $r \times r$ identity matrix (which follows from the identification conditions used in principal component analysis since the factors are only identified up to rotation). Notice that the dynamic equation for $\boldsymbol{f}_t$ implies that we are making the assumption that $\boldsymbol{x}_t^{GT}$ is $I(1)$ of dimension n, and $\boldsymbol{f}_t$ is $I(1)$ of dimension r. Later in this section the need for these assumptions will become clearer; the intuition behind them is that the factors and the change in unemployment, $R_t^{k,y}$, must be of the same order of integration.

Among others, Bai (2004) proves the consistency of the estimator of $I(1)$ factors by principal component analysis (PCA), under the assumptions of limited time and cross-sectional dependence and stationarity of the idiosyncratic components, $\boldsymbol{\varepsilon}_t$, and non-trivial contributions of the factors to the variance of $\boldsymbol{x}_t$.[4] We assume no cointegrating relationships among the factors. We further assume normality of the innovations for the same reasons outlined in Section 2.3.1.

The consistency of the two-step estimator has been originally proven in the stationary framework by Doz et al. (2011), and extended to the nonstationary case by Barigozzi and Luciani (2017).

In the first step, the factors ($\boldsymbol{f}_t$), the factor loadings ($\boldsymbol{\Lambda}$), and the covariance matrix of the idiosyncratic components ($\boldsymbol{\Psi}$) in model (2.3.3) are estimated by PCA as in Bai (2004). The matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are then replaced, in model (2.3.3), by their estimates $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}} = \mathrm{diag}\left(\hat{\psi}_{11}, \ldots, \hat{\psi}_{nn}\right)$ obtained in this first step. These estimates are kept fixed in the second step, because their high-dimensionality and associated curse of dimensionality complicates re-estimation by maximum likelihood. Moreover, restricting the covariance matrix of the idiosyncratic components $\boldsymbol{\Psi}$ to being diagonal is standard in the literature[5].

In order to make use of the auxiliary series to nowcast the unemployment, we stack together the measurement equations for $\boldsymbol{y}_t^k$ and $\boldsymbol{x}_t^{k,GT}$, respectively (2.3.1) and the first equation of (2.3.3) with $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ replaced, respectively, by $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\Psi}}$, and express them at the lowest frequency (in our case the monthly observation's frequency of $\boldsymbol{y}_t^k$). The transition equations for the RGB and survey error component in combination with the rotation scheme applied in the Dutch LFS hamper a formulation of the model on the high frequency. This means that $\boldsymbol{x}_t^{GT}$ needs to be first temporally aggregated from the high to the low frequency (either before or after the first step which estimates $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$). Since $\boldsymbol{x}_t^{GT}$ are the $I(1)$ weekly Google Trends, which are flow variables as they measure the proportion of queries made during each week, they are

[4] For the exact formulation we refer to Assumptions A-D in Bai (2004).

[5] The specification of the dynamic factor factor model with spherical idiosyncratic components is often called the "approximate" dynamic factor model. Doz et al. (2011) and Barigozzi and Luciani (2017) mention that misspecifications of this model arising from time or cross-sectional dependence of the idiosyncratic components, do not affect the consistency of the two-step estimator of the unobserved common factors, if n is large.

aggregated according to the following rule (Bańbura et al., 2013):

$$\boldsymbol{x}_{j,t}^{k,GT} = \sum_{i=1}^{j} \boldsymbol{x}_{t-k+i}^{GT}, \quad j = 1, \ldots, k, \quad t = k, 2k, \ldots, \quad k = \{4, 5\} \quad . \tag{2.3.4}$$

The aggregated $\boldsymbol{x}_{j,t}^{k,GT}$ are then rescaled (i.e., divided by their maximum value) in order to be bounded again between 0 and 100. The subscript j allows for real-time updating of the aggregated Google Trends in week j when new data become available. As such, this index indicates that we aggregate weeks 1 up to j. When $j = k$ we are at the end of the month, and we simply write $\boldsymbol{x}_t^{k,GT}$ to indicate the end-of-month aggregate value.

In order to get the final model, we also include a measurement equation for the univariate auxiliary series of the claimant counts, assuming that its state vector, $\theta_t^{k,CC}$, has the same composition as our population parameter $\theta_t^{k,y}$ (i.e., composed of a smooth trend and a seasonal component):

$$\begin{pmatrix} \boldsymbol{y}_t^k \\ x_t^{k,CC} \\ \boldsymbol{x}_t^{k,GT} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\imath}_5 \theta_t^{k,y} \\ \theta_t^{k,CC} \\ \hat{\boldsymbol{\Lambda}} \boldsymbol{f}_t^k \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t^k \\ 0 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_t^k \\ \varepsilon_t^{k,CC} \\ \boldsymbol{\varepsilon}_t^{k,GT} \end{pmatrix}, \tag{2.3.5}$$

$$\begin{pmatrix} \varepsilon_t^{k,CC} \\ \boldsymbol{\varepsilon}_t^{k,GT} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_{\varepsilon,CC}^2 & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Psi}} \end{bmatrix}\right),$$

$$\begin{pmatrix} \theta_t^{k,y} \\ \theta_t^{k,CC} \end{pmatrix} = \begin{pmatrix} L_t^{k,y} \\ L_t^{k,CC} \end{pmatrix} + \begin{pmatrix} S_t^{k,y} \\ S_t^{k,CC} \end{pmatrix}, \tag{2.3.6}$$

$$\begin{pmatrix} L_t^{k,y} \\ L_t^{k,CC} \end{pmatrix} = \begin{pmatrix} L_{t-1}^{k,y} \\ L_{t-1}^{k,CC} \end{pmatrix} + \begin{pmatrix} R_{t-1}^{k,y} \\ R_{t-1}^{k,CC} \end{pmatrix}, \tag{2.3.7}$$

$$\begin{pmatrix} R_t^{k,y} \\ R_t^{k,CC} \\ \boldsymbol{f}_t^k \end{pmatrix} = \begin{pmatrix} R_{t-1}^{k,y} \\ R_{t-1}^{k,CC} \\ \boldsymbol{f}_{t-1}^k \end{pmatrix} + \begin{pmatrix} \eta_{R,t}^{k,y} \\ \eta_{R,t}^{k,CC} \\ \boldsymbol{u}_t^k \end{pmatrix}, \tag{2.3.8}$$

$$\operatorname{cov}\begin{pmatrix} \eta_{R,t}^{k,y} \\ \eta_{R,t}^{k,CC} \\ \boldsymbol{u}_t^k \end{pmatrix} = \\ = \begin{bmatrix} \sigma_{R,y}^2 & \rho_{CC}\sigma_{R,y}\sigma_{R,CC} & \rho_{1,GT}\sigma_{R,y} & \cdots & \rho_{r,GT}\sigma_{R,y} \\ \rho_{CC}\sigma_{R,y}\sigma_{R,CC} & \sigma_{R,CC}^2 & 0 & \dots & 0 \\ \rho_{1,GT}\sigma_{R,y} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{r,GT}\sigma_{R,y} & 0 & 0 & \dots & 1 \end{bmatrix}. \tag{2.3.9}$$

The last equality allows the innovations of the trends' slopes, $R_t^{k,y}$ and $R_t^{k,CC}$, and of the factors of the Google Trends, to be correlated. Harvey and Chung (2000) show that there can be potential gains in precision, in terms of Mean Squared Error (MSE) of the Kalman filter estimators of $\theta_t^{k,y}$, $L_t^{k,y}$, and $R_t^{k,y}$, if the correlation parameters $|\rho|$ are large. Specifically, if $|\rho_{CC}| = 1$, then $\boldsymbol{y}_t^k$ and $x_t^{k,CC}$ have a common slope. This means that $\boldsymbol{y}_t^k$ and $x_t^{k,CC}$ are both $I(2)$, but there is a linear combination of their first differences which is stationary. Likewise, if $|\rho_{m,GT}| = 1$ then the m^{th} factor of the Google Trends and the change in unemployment, $R_t^{k,y}$, are cointegrated (i.e., they have the same source of error). This is why we need the elements of the vector in (2.3.8) to have the same order of integration, and it is via these correlation parameters that we exploit the auxiliary information.

The second step of the estimation procedure consists of estimating the remaining hyperparameters of the whole state space model (equations (2.3.5)-(2.3.9)) by maximum likelihood, and applying the Kalman filter to re-estimate $\boldsymbol{f}_t^k$ and to nowcast the variables of interest, $\theta_t^{k,y}$, $L_t^{k,y}$, and $R_t^{k,y}$, providing unemployment estimates in real-time before LFS data become available: $\hat{\theta}_{t|\Omega_t^-}^{k,y}$, $\hat{L}_{t|\Omega_t^-}^{k,y}$, and $\hat{R}_{t|\Omega_t^-}^{k,y}$ are the filtered nowcasts of, respectively, $\theta_t^{k,y}$, $L_t^{k,y}$, and $R_t^{k,y}$ based on the information set Ω_t^- available in month t. The information set in this case is $\Omega_t^- = \{\boldsymbol{x}_t^{k,GT}, \boldsymbol{y}_{t-1}^k, x_{t-1}^{k,CC}, \boldsymbol{x}_{t-1}^{k,GT}, \ldots\}$. Note that, contrary to Section 2.3.1, we now talk about "nowcast" instead of "forecast" of $\theta_t^{k,y}$ because a part of the data (the Google Trends) used in model (2.3.5)-(2.3.9) is now available in month t.

Some remarks are in order. First, although in Section 2.3.1 we mentioned that Statistics Netherlands publishes only $\hat{L}_t^{k,y}$ and $\hat{\theta}_t^{k,y}$ as official statistics for unemployment, we are also interested in the estimation/nowcast accuracy of $R_t^{k,y}$ since it is the state variable of the labour force model that is directly related to the auxiliary series.

Second, note that in model (2.3.3) we do not make use of the superscript k, meaning that the first step of the estimation can be performed on the high frequency (weekly in our empirical case) variables. Since in each week we can aggregate the weekly Google Trends to the monthly frequency, we can use the information available throughout the month to update the estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. If the correlations between the factors and the trend's slope of the target variable are large, this update should provide a more precise nowcast of $R_t^{k,y}$, $L_t^{k,y}$ and $\theta_t^{k,y}$.

Third, we allow the factors of the Google Trends data to be correlated with the change in unemployment and not with its level for two reasons: first, a smooth trend model is assumed for the population parameter, which means that the level of its trend does not have an innovation term. Second, it is reasonable to assume that people start looking for a job on the internet when they become unemployed, and hence their search behaviour should reflect the change in unemployment rather than its level[6].

[6] However, since in equation (2.3.7) $L_t^{k,y}$ depends on the *lagged* change in unemployment, $R_{t-1}^{k,y}$, it also depends, via the correlation parameters, on the *lagged* Google Trends' factors. This implies that the relationship between the *level* of unemployment and the Google Trends is not contemporaneous in our model specification. An alternative one that preserves the more intuitive contemporaneous relationship also between $L_t^{k,y}$ and $\boldsymbol{f}_t^k$, would be made of the following trends' transition equations:

$$\begin{pmatrix} L_t^{k,y} \\ R_t^{k,y} \\ \boldsymbol{f}_t^k \end{pmatrix} = \begin{pmatrix} L_{t-1}^{k,y} + R_{t-1}^{k,y} \\ R_{t-1}^{k,y} \\ \boldsymbol{f}_{t-1}^k \end{pmatrix} + \begin{bmatrix} 1 & \mathbf{0} \\ 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_r \end{bmatrix} \begin{pmatrix} \eta_t^{k,y} \\ \boldsymbol{u}_t^k \end{pmatrix},$$

with $\eta_t^{k,y}$ and $\boldsymbol{u}_t^k$ being correlated, and similarly for the trend's transition equations of the claimant count series.

Fourth, while our method to include auxiliary information in a state space model is based on the approach proposed by Harvey and Chung (2000), the factors of the high-dimensional auxiliary series could also be included as regressors in the observation equation for the labour force. However, in such a model, the main part of the trend, $L_t^{k,y}$, will be explained by the auxiliary series in the regression component. As a result, the filtered estimates for $L_t^{k,y}$ will contain a residual trend instead of the trend of the unemployment. Since the filtered trend estimates are the most important target variables in the official monthly publications of the labour force, this approach is not further investigated in this chapter[7]. Alternatively, the Google Trends' factors, extracted by PCA, could be treated as observed series and included in the model instead of $\boldsymbol{x}_t^{k,GT}$. This strategy would greatly simplify the model as its dimension would be reduced, but it may result in a possible loss in estimation accuracy of the factors, which is achieved by their second estimation with the Kalman filter (Doz et al., 2011)[8].

Finally, we refer the reader to Appendices 2.A.1, 2.A.2, and 2.A.3 for a detailed state space representation of the labour force model when, respectively, a univariate, a high-dimensional or both type of auxiliary series are included. We further refer to Appendices 2.A.2 for an illustration on how to include the lags of the factors and how to model their cycle or seasonality, within our proposed high-dimensional state space model.

2.4 Simulation study

We next conduct a Monte Carlo simulation study in order to elucidate to what extent our proposed method can provide gains in the nowcast accuracy of the unobserved components of interest. For this purpose, we consider a simpler model than the one used for the labour force survey. Here y_t^k is univariate

[7] Moreover, this alternative approach would allow us to achieve better estimates and nowcasts of the observed series, $\boldsymbol{y}_t^k$, rather than the unobserved components of interest, which is instead what we are interested in.

[8] Technically it is the Kalman smoother that achieves the additional estimation efficiency, but in the last period of the sample, which is what we are interested in when nowcasting, the Kalman filter and smoother estimates are the same.

following a smooth trend model, and $\boldsymbol{x}_t^k$ represents the (100×1)-dimensional auxiliary series with one common factor ($r = 1$).

$$
\begin{aligned}
\begin{pmatrix} y_t^k \\ \boldsymbol{x}_t^k \end{pmatrix} &= \begin{pmatrix} L_t^k \\ \Lambda f_t^k \end{pmatrix} + \begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix}, \\
L_t^k &= L_{t-1}^k + R_{t-1}^k, \\
\begin{pmatrix} R_t^k \\ f_t^k \end{pmatrix} &= \begin{pmatrix} R_{t-1}^k \\ f_{t-1}^k \end{pmatrix} + \begin{pmatrix} \eta_{R,t}^k \\ u_t^k \end{pmatrix}, \quad \begin{pmatrix} \eta_{R,t}^k \\ u_t^k \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).
\end{aligned}
$$

We allow the slope's and factor's innovations to be correlated, and we investigate the performance of the method for increasing values of the correlation parameter $\rho \in [0, 0.2, 0.4, 0.6, 0.8, 0.9, 0.99]$. The auxiliary variable $\boldsymbol{x}_t^k$ has the same frequency as y_t^k and it is assumed that all $\boldsymbol{x}_t^k$ are released at the same time without publication delays. The nowcast is done concurrently, i.e. in real-time based on a recursive scheme. This means that at each time point of the out-of-sample period, the hyperparameters of the model are re-estimated by maximum likelihood, extending the same used up to that period. This is done in the third part of the sample, always assuming that y_t^k is not available at time t, contrary to $\boldsymbol{x}_t^k$. This implies that the available data set in period t equals $\Omega_t^- = \{\boldsymbol{x}_t^k, y_{t-1}^k, \boldsymbol{x}_{t-1}^k, y_{t-2}^k, \ldots\}$. The sample size is $T = 150$ and the number of simulations is $n_{\text{sim}} = 500$.

We consider three specifications for the idiosyncratic components and the factor loadings:

1. Homoskedastic idiosyncratic components and dense loadings:

$$
\begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N\left(\mathbf{0}, 0.5\boldsymbol{I}_{n+1}\right), \quad \Lambda \sim U\left(0, 1\right).
$$

2. Homoskedastic idiosyncratic components and sparse loadings. The first half of the elements in the loadings are set equal to zero. This specification reflects the likely empirical case that some of the Google Trends

are not related to the change in unemployment:

$$\begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N\left(\mathbf{0}, 0.5\boldsymbol{I}_{n+1}\right),$$

$$\Lambda = \left(\Lambda_0', \Lambda_1'\right)', \underset{50\times 1}{\Lambda_0} = \mathbf{0}, \underset{50\times 1}{\Lambda_1} \sim U\left(0,1\right).$$

3. Heteroskedastic idiosyncratic components and dense loadings. The homoskedasticity assumption is here relaxed, again as not being realistic for the job search terms:

$$\begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} 0.5 & \mathbf{0}' \\ \mathbf{0} & \text{diag}(H) \end{pmatrix}\right),$$

$$H \sim U(0.5, 10), \quad \Lambda \sim U\left(0,1\right).$$

Let $\boldsymbol{\alpha}_t^k = \left(L_t^k, R_t^k, f_t^k\right)'$ denote the vector of state variables and $\hat{\boldsymbol{\alpha}}_{t|\Omega_t^-}^k$ its estimates based on the information available at time t. The results from the Monte Carlo simulations are shown in Table 2.2. We always report the MSFE, together with its variance and bias components, of the Kalman filter estimator of $\boldsymbol{\alpha}_t^k$, relative to the same measures calculated from the model that does not include the auxiliary series $\boldsymbol{x}_t^k$. Recall that the latter comes down to making one-step-ahead predictions.

$$\text{MSFE}(\hat{\boldsymbol{\alpha}}_{t|\Omega_t^-}^k) = \frac{1}{h}\sum_{t=T-h+1}^{T} \frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}} \left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right)\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-} - \boldsymbol{\alpha}_{jt}\right)',$$

$$
\begin{aligned}
&\mathrm{var}(\hat{\boldsymbol{\alpha}}^k_{t|\Omega_t^-}) = \\
&= \frac{1}{h}\sum_{t=T-h+1}^{T}\left(\frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}}\left(\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-}-\boldsymbol{\alpha}_{jt}\right)-\frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}}\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-}-\boldsymbol{\alpha}_{jt}\right)\right)\right.\\
&\qquad\left.\times\left(\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-}-\boldsymbol{\alpha}_{jt}\right)-\frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}}\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-}-\boldsymbol{\alpha}_{jt}\right)\right)'\right),
\end{aligned}
$$

$$
\begin{aligned}
&\mathrm{bias}^2(\hat{\boldsymbol{\alpha}}^k_{t|\Omega_t^-}) = \\
&= \frac{1}{h}\sum_{t=T-h+1}^{T}\left(\frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}}\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-}-\boldsymbol{\alpha}_{jt}\right)\right)\left(\frac{1}{n_{\text{sim}}}\sum_{j=1}^{n_{\text{sim}}}\left(\hat{\boldsymbol{\alpha}}_{jt|\Omega_t^-}-\boldsymbol{\alpha}_{jt}\right)\right)',
\end{aligned}
$$

where h is the size of the out-of-sample period.

In every setting, both the bias and the variance components of the MSFE tend to decrease with the magnitude of the correlation parameter. The improvement is more pronounced for the slope rather than the level of the trend. For the largest value of the correlation, with respect to the model which does not include auxiliary information, the gain in MSFE for the level and the slope is, respectively, of around 25% and 75%. Moreover, for low values of ρ, the MSFE does not deteriorate with respect to the benchmark model. This implies that our proposed method is robust to the inclusion of auxiliary information that does not have predictive power for the state variables of interest. In Appendix 2.C we report and examine additional simulation results with non-Gaussian idiosyncratic components, and draw the same conclusions discussed above for the MSFE and the variance of the state variables' nowcasts. The bias instead worsens while deviating from Gaussianity, but it does not affect the MSFE as it only accounts for a small part of the latter measure. We therefore conclude that the performance of our method is overall robust to deviations from Gaussianity of the idiosyncratic components.

The decision to focus the simulation study on the nowcast (rather than the in-sample) performance of our method, is motivated by the fact that the added

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.99$
	Homoskedastic idiosyncratic components and dense loadings						
MSFE($\hat{L}^k_{t\|\Omega_t^-}$)	1.030	1.024	1.006	0.971	0.901	0.837	0.718
var($\hat{L}^k_{t\|\Omega_t^-}$)	1.031	1.025	1.007	0.971	0.901	0.837	0.718
bias^2($\hat{L}^k_{t\|\Omega_t^-}$)	0.775	0.767	0.756	0.733	0.692	0.659	0.567
MSFE($\hat{R}^k_{t\|\Omega_t^-}$)	1.044	1.017	0.941	0.806	0.588	0.427	0.198
var($\hat{R}^k_{t\|\Omega_t^-}$)	1.045	1.018	0.942	0.807	0.589	0.427	0.198
bias^2($\hat{R}^k_{t\|\Omega_t^-}$)	0.650	0.633	0.583	0.492	0.350	0.252	0.122
	Homoskedastic idiosyncratic components and sparse loadings						
MSFE($\hat{L}^k_{t\|\Omega_t^-}$)	1.031	1.026	1.011	0.981	0.920	0.862	0.744
var($\hat{L}^k_{t\|\Omega_t^-}$)	1.031	1.026	1.012	0.981	0.920	0.862	0.745
bias^2($\hat{L}^k_{t\|\Omega_t^-}$)	0.784	0.776	0.762	0.737	0.695	0.655	0.582
MSFE($\hat{R}^k_{t\|\Omega_t^-}$)	1.044	1.019	0.946	0.817	0.605	0.446	0.208
var($\hat{R}^k_{t\|\Omega_t^-}$)	1.045	1.020	0.947	0.817	0.606	0.446	0.209
bias^2($\hat{R}^k_{t\|\Omega_t^-}$)	0.656	0.639	0.586	0.492	0.347	0.243	0.104
	Heteroskedastic idiosyncratic components and dense loadings						
MSFE($\hat{L}^k_{t\|\Omega_t^-}$)	1.036	1.032	1.019	0.994	0.945	0.901	0.823
var($\hat{L}^k_{t\|\Omega_t^-}$)	1.037	1.032	1.020	0.995	0.946	0.902	0.823
bias^2($\hat{L}^k_{t\|\Omega_t^-}$)	0.707	0.645	0.579	0.521	0.484	0.483	0.543
MSFE($\hat{R}^k_{t\|\Omega_t^-}$)	1.049	1.027	0.960	0.840	0.644	0.499	0.299
var($\hat{R}^k_{t\|\Omega_t^-}$)	1.049	1.028	0.961	0.841	0.645	0.500	0.299
bias^2($\hat{R}^k_{t\|\Omega_t^-}$)	0.805	0.697	0.556	0.397	0.230	0.161	0.237

Table 2.2: Simulation results from the three settings described in Section 2.4. The values are reported relative to the respective measures calculated from the model that does not include the auxiliary series; values < 1 are in favour of our method. $n_{\text{sim}} = 500$.

value of the Google Trends over the claimant counts is their real-time availability, which can be used to nowcast the unemployment. Nonetheless, for completeness, in the empirical application of the next section we report the results also for the in-sample performance of our method.

2.5 Application to Dutch unemployment nowcasting

In this section we present and discuss the results of the empirical application of our method to nowcasting the Dutch unemployment using the auxiliary series of claimant counts and Google Trends related to job-search and economic uncertainty.

As explained in Section 2.3.2, the Google series used in the model must be $I(1)$. We therefore test for nonstationarity in the Google Trends with the Elliott et al. (1996) augmented Dickey-Fuller (ADF) test, including a constant and a linear trend. We control for the false discovery rate as in Moon and Perron (2012), who employ a moving block bootstrap approach that accounts for time and cross-sectional dependence among the units in the panel.

Before proceeding with the estimation of the model by only including the Google Trends that resulted as being $I(1)$ from the multiple hypotheses testing, we carry out an additional selection of the $I(1)$ Google Trends by "targeting" them as explained and motivated below.

Bai and Ng (2008) point out that having more data to extract factors from is not always better. In particular, if series are added that have loadings of zero and are thus not influenced by the factors, these will make the estimation of factors and loadings by PCA deteriorate, as PCA assigns a non-zero weight to each series in calculating the estimated factor as a weighted average. Bai and Ng (2008) recommend a simple strategy to filter out irrelevant series (in our case Google search terms) and improve the estimation of the factors, which they call "targeting the predictors". In this case an initial regression of the series of interest is performed on the high-dimensional input series to determine which series are (ir)relevant. The series that are found to be irrelevant are discarded and only the ones that are found to be relevant are kept to estimate the factors and loadings from. In particular, they recommend the use of the elastic net (Hastie & Zou, 2005), which is a penalized regression technique that performs estimation and variable selection at the same time by setting the coefficients of the irrelevant variables to 0 exactly. After performing the elastic net estimation, only the variables with non-zero coefficients are then kept. As we do not observe our series of interest directly, we need to adapt their procedure to our setting. To do so we approximate the unobserved unemployment by its estimation from the labour force model without auxiliary series. Specifically, we regress the differenced estimated change in unemployment from the labour force model without auxiliary series, $\Delta\hat{R}_t^{k,y}$, on the differenced $I(1)$ Google Trends using the elastic net penalized regression method, which solves the following minimization problem:

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{2T} \sum_{t=1}^{T} \left(\Delta\hat{R}_t^{k,y} - \boldsymbol{\beta}' \Delta\boldsymbol{x}_t^{k,GT} \right)^2 + \lambda P_\alpha \left(\boldsymbol{\beta} \right) \right],$$

where

$$P_\alpha \left(\boldsymbol{\beta} \right) = (1-\alpha) \frac{1}{2} ||\boldsymbol{\beta}||_2^2 + \alpha ||\boldsymbol{\beta}||_1.$$

The tuning parameters λ and α are selected from a two-dimensional grid in order to minimize the Schwarz (1978) Bayesian information criterion (BIC). Notice that performing the penalized regression on the differenced (and therefore stationary) data, also allows us to avoid the inclusion in the model of Google Trends that have spurious relations with the change in unemployment.

We consider estimating the final model both with all Google Trends included and with only the selected Google Trends included, thereby allowing us to assess the empirical effects of targeting. The final number of nonstationary Google Trends included in the model, n, may differ depending on whether we use the weekly Google Trends aggregated to the monthly frequency according to equation (2.3.4), or the monthly Google Trends. Whenever we apply PCA, the Google Trends are first differenced and standardized.

We further need to make sure that the stationarity assumption of the idiosyncratic components is maintained. Therefore, after having estimated the factors by PCA in model (2.3.3), we test which of the idiosyncratic components $\boldsymbol{\varepsilon}_t$ are $I(1)$ with an ADF test without deterministic components, by controlling for multiple hypotheses testing as in Moon and Perron (2012). The $I(1)$ idiosyncratic components are modelled as state variables in (2.3.5), with the following transition equation:

$$\varepsilon_t^k = \varepsilon_{t-1}^k + \boldsymbol{\xi}_t^k,$$

with usual normality assumptions on the $\boldsymbol{\xi}_t^k$. The covariance matrix of the idiosyncratic components $\boldsymbol{\Psi}$ is therefore estimated on the levels of the $I(0)$ idiosyncratic components and the first differences of the $I(1)$ idiosyncratic components. Appendix 2.A.2 provides a toy example that elucidates the estimation procedure.

Finally, we notice that although the first step of the two-step estimation procedure is meant to avoid estimating $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$ by maximum likelihood (since they are large matrices), this pre-estimation may affect the explanatory power of the Google Trends. We here propose two different ways to obtain (possibly) more accurate estimates of these two matrices:

- In Section 2.3.2 we mention that the first step of the two-step estimator, which estimates $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$ by PCA, can be carried out on the weekly Google Trends (which are therefore aggregated to the monthly frequency after the first step). Since the sample size of the high frequency data is larger, using weekly Google Trends might improve the estimation accuracy of $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$.

- Doz et al. (2011) argue that from the Kalman filter estimates of the factors, it is possible to re-estimate $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$ (by least squares), which in turn can be used to re-estimate the factors, and so on. This iterative procedure is equivalent to the Expectation–Maximization (EM) algorithm, which increases the likelihood at each step and therefore converges to the maximum likelihood solution. Notice that since the Kalman filter can (in our setting) only provide monthly estimates, the iterative estimation is done on the low-frequency Google Trends.

Later in this section we check how sensitive our empirical results are to the different estimates of $\boldsymbol{\Psi}$ and $\boldsymbol{\Lambda}$. For the second type of estimation method discussed above, we only perform one additional iteration of the two-step procedure due to its computational burden.

We present empirical results for the in-sample estimates and out-of-sample forecasts. With the in-sample estimates we evaluate to what extent the auxiliary series improve the precision of the published monthly unemployment estimates after finalizing the data collection. With the out-of-sample forecasts we evaluate to which extent the auxiliary series improve the precision of provisional estimates in a nowcast procedure during the period of data collection. We always estimate four different models: the labour force model without auxiliary series (baseline), the labour force model with auxiliary series of claimant counts (CC), of Google Trends (GT) and of both (CC & GT). We compare the latter three models to the baseline one with the in-sample and out-of-sample exercises. The period considered for the estimation starts in January 2004 and ends in May 2019 ($T = 185$ months). The out-of-sample nowcasts are conducted in real-time (concurrently) in the last three years of the sample based on a recursive scheme: each week or month, depending on whether we use weekly or monthly Google Trends, the model, including its hyperparameters, is re-estimated on the enlarged sample now extended by the latest observations, while assuming that the current observations for the unemployed labour force and the claimant counts are missing. Analogously, when the Google Trends are first targeted with the elastic net, the targeting is re-executed in each week or month of the out-of-sample period on the updated sample.[9]

We define the measure of in-sample estimation accuracy $\widehat{\mathrm{MSE}}(\hat{\boldsymbol{\alpha}}^k_{t|\Omega_t}) = \frac{1}{T-d}\sum_{t=d+1}^{T} \hat{\boldsymbol{P}}^k_{t|\Omega_t}$, where $\hat{\boldsymbol{\alpha}}^k_{t|\Omega_t}$ is the vector of Kalman filter estimates of the state variables, $\hat{\boldsymbol{P}}^k_{t|\Omega_t}$ is its estimated covariance matrix in month t, and d is the number of state variables that are needed to estimated the labour force model without auxiliary series, and that need a diffuse initialization for their estimation ($d = 17$). The measure of nowcast accuracy, $\widehat{\mathrm{MSFE}}(\hat{\boldsymbol{\alpha}}^k_{t|\Omega_t^-}) = \frac{1}{h}\sum_{t=T-h+1}^{T} \hat{\boldsymbol{P}}^k_{t|\Omega_t^-}$, is the average of the nowcasted covariance matrices in the h prediction months. When

[9] The elastic net is nowadays implemented very efficiently in standard statistical software, so the targeting step is not computationally expensive.

weekly Google Trends are used, $\hat{\boldsymbol{P}}^k_{t|\Omega_t^-} = \frac{1}{k}\sum_{j=1}^k \hat{\boldsymbol{P}}^k_{j|\Omega^-_{j,t}}$, where $\hat{\boldsymbol{P}}^k_{j|\Omega^-_{j,t}}$ is the nowcasted covariance matrix for the prediction in week j of month t, and $\Omega^-_{j,t} = \{\boldsymbol{x}^{k,GT}_{j,t}, \boldsymbol{y}^k_{t-1}, x^{k,CC}_{t-1}, \boldsymbol{x}^{k,GT}_{t-1}, \ldots\}$ is in this case the available information set in week j of month t. This is because the nowcast is done recursively throughout the weeks of the out-of-sample period. We always report the relative $\widehat{\text{MS(F)E}}$ with respect to the baseline model; values lower than one are in favour of our method. We note that nowcasting under the baseline model without auxiliary series and the baseline model extended with claimant counts comes down to making one-step-ahead predictions. Therefore, it should be kept in mind that, although for simplicity we always talk about “nowcast” performance of the different models, the claimant counts can only improve the “one-step-ahead forecast” performance of the models that include them. Expressions for $\hat{\boldsymbol{\alpha}}^k_{t|\Omega_t}$, $\hat{\boldsymbol{\alpha}}^k_{t|\Omega_t^-}$ and their covariance matrices, $\hat{\boldsymbol{P}}^k_{t|\Omega_t}$ and $\hat{\boldsymbol{P}}^k_{t|\Omega_t^-}$, are given by the standard Kalman filter recursions, see e.g. Durbin and Koopman (2012, Chapter 4).

The initial values of the hyperparameters for the maximum likelihood estimation are equal to the estimates for the labour force model obtained in van den Brakel and Krieg (2015). We use a diffuse initialisation of the Kalman filter for all the state variables except for the 13 state variables that define the autocorrelation structure of the survey errors, for which we use the exact initialisation of Bollineni-Balabay et al. (2017).

We use the three panel information criteria proposed by Bai and Ng (2002) which we indicate, as in Bai and Ng (2002), with IC_1, IC_2 and IC_3, in order to choose how many factors of the Google Trends to include in the model[10]. When the Google Trends are targeted with the elastic net, the information criteria suggest to include one or two factors. In the empirical analysis we check the sensitivity of the results with respect to these two different numbers of factors included in the model.

We employ a Wilks (1938) likelihood ratio (LR) test to assess whether the correlation parameters are significantly different from zero, and hence adding the

[10] In this chapter, if for instance the information criterion IC_1 suggests to include 2 factors, we indicate as $IC_1 = 2$.

auxiliary information might yield a significant improvement from the baseline model. Specifically, we indicate with $\rho_{CC} = 0$, $\rho_{1,GT} = 0$ and $\rho_{2,GT} = 0$ the null hypotheses for the individual insignificance of the correlation parameter with, respectively, the claimant counts, and the first and second factor (when present) of the Google Trends. With $\boldsymbol{\rho}_{GT} = \mathbf{0}$ and $\boldsymbol{\rho} = \mathbf{0}$ we instead indicate the null hypotheses for the joint insignificance of, respectively, the correlations with the Google Trends' factors, and all correlation parameters. If the true distribution of the error terms is non-Gaussian, the LR test, based on the QML estimates, does not generally retain, under the null hypothesis, an asymptotic χ^2 distribution with degrees of freedom equal to the number of restrictions. One exception is when the covariance matrix of the error terms from a regression involving observed variables, is replaced by a consistent estimator prior to the maximization of the log-likelihood (Gourieroux & Monfort, 1993). In our case, if the idiosyncratic components of the Google Trends, $\varepsilon_t^{k,GT}$, are the only error terms not being normally-distributed, we may fall into this exception. The covariance matrix $\boldsymbol{\Psi}$ is indeed replaced, for the maximization of the log-likelihood, by its consistent PCA estimator obtained in the first step of the two-step estimation procedure. Nonetheless, in the setting of Gourieroux and Monfort (1993) the regressors are observed, whereas in our case the latter are the unobserved factors. Consequently, it is not trivial to assess whether our model specification indeed falls into the above-mentioned exception. A formal proof for this is beyond the scope of this chapter, but in Appendix 2.C we conduct a simulation study in order to obtain the finite-sample probability density of the LR test under misspecifications of the distribution of the idiosyncratic components. We conclude that the distribution of the LR test is not affected by these misspecifications. At the end of this section we show that that there is no evidence that the error terms other than $\varepsilon_t^{k,GT}$, are not normally-distributed. We should therefore be able perform inference based on the usual asymptotic distribution of the LR test.

Table 2.3 reports the estimated hyperparameters for the four models, as well as the respective value for the maximized log-likelihood, the relative measures of in and out-of-sample performance, and the p-values from the LR tests, when the monthly Google Trends are used.

The maximum likelihood estimates for the standard error of the seasonal com-

			$n = 162, IC_1 = 3, IC_2 = 1, IC_3 = 10$		Targeted GT, $n = 39, IC_1 = 2, IC_2 = 1, IC_3 = 2$				
			$r = 1$		$r = 1$			$r = 2$	
	LF	CC	GT	CC & GT	GT	CC & GT	CC & GT, all corr.	GT	CC & GT
$\hat{\sigma}_{R,y}$	2082.652	2776.030	1995.917	2704.918	3036.281	2608.394	3447.973	3587.985	3002.947
$\hat{\sigma}_{\omega,y}$	0.020	0.020	0.023	0.078	0.013	0.011	0.022	0.054	0.010
$\hat{\sigma}_{\lambda}$	3841.035	3883.658	3592.394	3715.303	3740.097	3740.748	3115.596	3670.943	3709.361
$\hat{\sigma}_{\nu_1}$	1.140	1.151	1.181	1.146	1.155	1.142	1.205	1.155	1.198
$\hat{\sigma}_{\nu_2}$	1.291	1.300	1.270	1.359	1.276	1.304	1.378	1.281	1.263
$\hat{\sigma}_{\nu_3}$	1.188	1.181	1.201	1.211	1.188	1.196	1.117	1.224	1.200
$\hat{\sigma}_{\nu_4}$	1.240	1.247	1.241	1.224	1.241	1.252	1.356	1.286	1.243
$\hat{\sigma}_{\nu_5}$	1.223	1.228	1.236	1.260	1.221	1.239	1.358	1.254	1.247
$\hat{\delta}$	0.384	0.381	0.378	0.395	0.377	0.384	0.390	0.383	0.384
$\hat{\sigma}_{R,CC}$		3490.261		3515.222		3503.077	3982.583		3979.232
$\hat{\sigma}_{\omega,CC}$		0.020		0.020		0.021	0.016		0.020
$\hat{\sigma}_{\varepsilon,CC}$		1318.691		1310.108		1309.136	1181.291		1052.729
$\hat{\rho}_{CC}$		0.918		0.913		0.803	0.935		0.755
$\hat{\rho}_{1,GT}$			-0.200	-0.003	-0.899	-0.509	-0.250	-0.785	-0.381
$\hat{\rho}_{2,GT}$								-0.591	-0.456
$\hat{\rho}_{1,CC,GT}$							-0.093		
$\widehat{\text{MSE}}(\hat{L}^{k,y}_{t\vert\Omega_t})$		0.868	1.003	0.863	0.919	0.796	0.895	0.861	0.849
$\widehat{\text{MSE}}(\hat{R}^{k,y}_{t\vert\Omega_t})$		0.878	0.916	0.849	0.655	0.618	1.112	0.485	0.702
$\widehat{\text{MSE}}(\hat{\theta}^{k,y}_{t\vert\Omega_t})$		0.889	1.009	0.888	0.941	0.835	0.916	0.899	0.881
$\widehat{\text{MSFE}}(\hat{L}^{k,y}_{t\vert\Omega_t^-})$		0.818	0.951	0.853	0.875	0.766	0.786	0.935	0.889
$\widehat{\text{MSFE}}(\hat{R}^{k,y}_{t\vert\Omega_t^-})$		0.983	0.878	0.981	0.705	0.801	0.755	0.839	0.869
$\widehat{\text{MSFE}}(\hat{\theta}^{k,y}_{t\vert\Omega_t^-})$		0.827	0.956	0.860	0.886	0.779	0.796	0.942	0.899
log-likelihood	-10160.378	-11835.779	-44726.153	-46392.985	-18712.139	-20379.780	-20384.560	-18719.515	-20388.495
		p-value from the LR test							
$H_0 : \rho_{CC} = 0$		0.002		0.000		0.001	0.000		0.000
$H_0 : \rho_{1,GT} = 0$			0.470	0.830	0.001	0.025	0.028	0.000	0.082
$H_0 : \rho_{2,GT} = 0$								0.014	0.014
$H_0 : \boldsymbol{\rho}_{GT} = \mathbf{0}$								0.000	0.017
$H_0 : \rho_{1,CC,GT} = 0$							0.470		
$H_0 : \boldsymbol{\rho} = \mathbf{0}$				0.001		0.000	0.000		0.000

Table 2.3: Estimation and nowcast results for the labour force model with and without auxiliary series. The auxiliary series are the claimant counts and the monthly Google Trends about job-search and economic uncertainty. The number of Google Trends and the number of their factors included in the model are denoted with n and r, respectively. The abbreviation "all corr." denotes that the correlation between the claimant counts and the Google Trends is also estimated. "Targeted GT" indicates that the Google Trends have been targeted with the elastic net before including them in the model.

ponents' disturbance terms tend to zero, indicating that the seasonal effects are time invariant.

Recall from equation (2.3.2) that the variances of the scaled sampling errors, $\sigma^2_{\nu_j}$, should take values close to one. Their estimates are divided by (1 - $\hat{\delta}^2$) and are always slightly larger than one, which is an indication that the variance estimates of the GREG estimates, used to scale the sampling errors in equation (2.3.2), somewhat underestimate the real variance of the GREG estimates.

The correlation with the claimant counts is estimated to be above 0.9, and remains large and significant when including the Google Trends. Similar conclusions can be drawn for the correlations with the Google Trends' factors, when the Google Trends are targeted with the elastic net, and 39 of them are included in the model. When the additional targeting is not applied, and the 162 $I(1)$ Google Trends are directly included in the model, the correlation parameter with the first factor of the Google Trends is instead always small and insignificant (in this setting we do not include more than one factor). Moreover, for the same number of factors, targeting the Google Trends always yields a better performance in terms of estimation and nowcast accuracy of the state variables of interest, with respect to not targeting them. For this reason, we focus the remaining analysis of the empirical results only on the targeted Google Trends.

The best results in terms of both estimation and nowcast accuracy of all the state variables, is achieved by the CC & GT model with one factor, yielding a gain of, respectively, around 40% and 20% for $\hat{R}^{k,y}_{t|\Omega_t^-}$, and around 20% and 25% for both $\hat{L}^{k,y}_{t|\Omega_t^-}$ and $\hat{\theta}^{k,y}_{t|\Omega_t^-}$, with respect to the baseline model. Note that this implies that the above-mentioned model outperforms also the model that contains only the claimant counts as auxiliary series. In general, the models with Google Trends tend to achieve a better estimation and nowcast of the change in unemployment, $R^{k,y}_t$, rather than the other two state variables, with respect to the models that include the claimant counts.

Including two factors instead of one clearly increases the complexity of the model, which is reflected in smaller accuracy gains (in the CC & GT model probably also due to the decreased magnitude of the correlation parameter with the claimant counts), especially for the nowcast of the state variables, with respect to including only one factor. Nonetheless, the correlations with both factors are individually and jointly significantly different from zero, indicating that both factors bring additional information about the Dutch unemployment.

Notice that in general all the relative measures of accuracy are below one, indicating that both the claimant counts and the Google Trends improve the estimation and nowcast accuracy of the unemployment and its change. Even when the Google Trends are not targeted and their factor is not significantly related to the unemployment, the measures are never drastically above one, meaning that our method tends to ignore auxiliary series that are not related to the target variable.

Finally, when we specified the covariance matrix (2.3.9) in Section 2.3.2, we did not let the claimant counts and the Google Trends be correlated because our goal is to improve the estimation/nowcast accuracy of the unobserved components of the labour force series, not of the claimant counts nor the Google Trends. Nonetheless, if the state variables of equation (2.3.8) are all cointegrated (i.e. the correlation parameters are all equal to one) a more efficient estimation method would be to only estimate the variance of their common source of error. We therefore estimate the CC & GT model with one factor, when all series are correlated. We call this model "CC & GT all corr.". Table 2.3 reports the empirical results also for this model. Although the nowcast accuracy is similar to the same model without the additional correlation between the claimant counts and the Google Trends (which we indicate as $\rho_{1,CC,GT}$), the in-sample accuracy deteriorates (even with respect to the baseline model), and $\rho_{1,CC,GT}$ is not significantly different from zero. We therefore conclude that the specification of the covariance matrix (2.3.9) is appropriate.

In Table 2.4 we report the empirical results for the GT and CC & GT models which employ the targeted Google Trends observed at the weekly frequency, and aggregated to the monthly frequency according to equation (2.3.4) in order to include them in the models. In this case we still look at the sensitivity of the results with respect to the number of factors included in the model, but also with respect to the two additional methods for the estimation of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ discussed at the beginning of this section.

The measures of accuracy are again broadly lower than one, but the gains are not as large as observed for the monthly Google Trends. Including two factors improves the accuracy in the GT model, but not in the CC & GT model, except for a more precise nowcast of $R_t^{k,y}$. The correlation parameter with the claimant counts remains large and significant. On the contrary, the correlation parameter with the first factor of the Google Trends is not significantly different from zero, and there is a weak evidence for the second factor being significantly related to the change in unemployment. For this reason we continue the analysis by considering two factors in the model.

Estimating $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ on the weekly Google Trends improves the measures of accuracy only for the CC & GT model, and not for the GT model. An additional iteration of the two step estimator, in order to obtain more accurate estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, achieves instead better nowcasts for both the GT and the CC & GT models (and also better in-sample estimates for the latter model), and a similar performance to the models which employ the monthly Google trends and include two factors. Notice that the values of the log-likelihood for these two models increased with respect to the same model specifications that use the original two-step estimation (without the additional iteration). The latter result, as pointed out in the explanation of the iterated estimation of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ at the beginning of this section, is to be expected. Despite the above-mentioned improvements in estimation/nowcast accuracy, the correlation parameters with the Google Trends' factors are always insignificant. The aggregation of the Google Trends from the weekly to the monthly frequency yields time series that are more noisy with respect to the Google Trends that are directly observed at the monthly frequency, and detecting significant results therefore becomes harder.

Finally, even though weekly Google Trends allow us to perform the monthly nowcasts on a weekly basis, we notice that, in general, the precision of the nowcast does not monotonically improve with the number of weeks. If the high-dimensional state space model could be expressed and estimated at the highest frequency, the weekly gains in nowcast accuracy could be more evident. Nonetheless, we are limited by the transition equations for the RGB and the survey errors, to estimate the model at the monthly frequency.

Figures 2.2-2.4 compare the point nowcasts, respectively, of the change in unemployment, its trend, and the population parameter, obtained with the baseline, the CC, the GT and the CC & GT models which employ monthly Google Trends and include two of their factors. From the first graph, it is evident that the models including claimant counts tend to deviate from the baseline model. The latter, on the contrary, gives similar results as those of the GT model. The point nowcasts of $L_t^{k,y}$ and $\theta_t^{k,y}$ are more similar throughout the model specifications, with a slight and positive difference between the models that include the Google Trends and the ones that do not, at the beginning of the out-of-sample period.

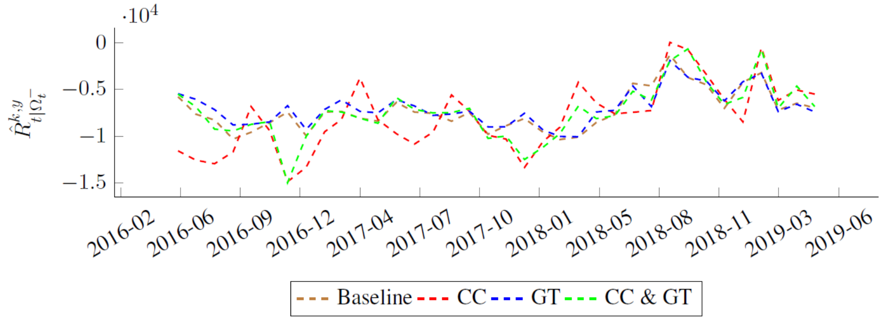


Figure 2.2: Nowcast of $R_t^{k,y}$ with the labour force models. The results for the GT and the CC & GT models refer to setting where the monthly Google Trends are used, and two of their factors are included in the model.

Figures 2.D.1 and 2.D.2 show the selection frequency of, respectively, the monthly and weekly Google Trends in the out-of-sample period. Some of the most selected search terms in both cases are: werklozen (unemployed people), baan zoeken (job search), curriculum vitae voorbeeld (curriculum vitae example), ww uitkering (unemployment benefits), ww aanvragen (to request unemployment benefits), resume, tijdelijk werk (temporary job), huizenmarkt zeepbel (housing market bubble). Notice that the latter term (as well as “economische crisis” (economic crisis) or “failliet” (bankrupt), which are also frequently selected monthly Google Trends) is of economic uncertainty nature, rather than being job-search related. Other investigations not reported here only used the latter type of search terms, and did not find them to have explanatory power for Dutch unemployment, which is now instead significantly improved by the inclusion of search terms related to economic uncertainty.

The results of the empirical analysis can be summarized as follows. Target-

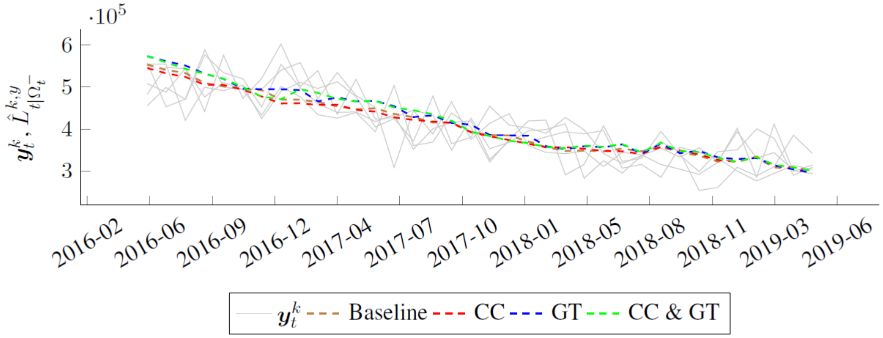


Figure 2.3: Nowcast of $L_t^{k,y}$ with the labour force models, compared to the five waves of the unemployed labour force. The results for the GT and the CC & GT models refer to the setting where the monthly Google Trends are used, and two of their factors are included in the model.

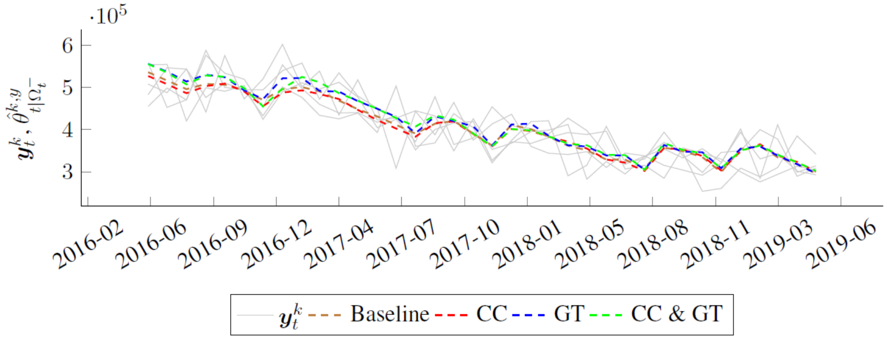


Figure 2.4: Nowcast of $\theta_t^{k,y}$ with the labour force models, compared to the five waves of the unemployed labour force. The results for the GT and the CC & GT models refer to the setting where the monthly Google Trends are used, and two of their factors are included in the model.

ing the Google Trends improves its explanatory power for the Dutch unemployment. Monthly Google Trends significantly improve the estimation and nowcast accuracy of Dutch unemployment and its change, with both one and two factors. The largest gains are obtained when both the claimant count and the Google Trends are included, and considering only one factor for the latter series. When two factors are considered, the gains are smaller but both factors seem to be significantly related to the change in unemployment, indicating that both of them should be included in the model in order to exploit all the information that the Google Trends give about the target variable. The sensitivity to the number of factors is somewhat similar for the weekly Google Trends, although there is only weak evidence for their second factor to have a significant relation with the change in unemployment. The weekly Google Trends are less informative about the Dutch unemployment, yielding in general less improvement in estimation and nowcast accuracy, with respect to the monthly Google Trends. The contributions of the two types of Google Trends are comparable only when the two-step estimator is additionally re-iterated for the weekly Google Trends (in order to obtain more precise estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$). This result suggests that iterating the two-step estimation can improve the explanatory power of the Google Trends, and that the latter series are sensitive to the estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. Improvements are, instead, not always present when $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are estimated on the weekly data. In general, the claimant counts mainly have a positive impact on the estimation and nowcast accuracy of $\theta_t^{k,y}$ and $L_t^{k,y}$, whereas the Google Trends affect $R_t^{k,y}$. The point nowcasts of the latter state variable are more sensitive to the type of auxiliary series included, than those of $\theta_t^{k,y}$ and $L_t^{k,y}$.

The assumptions of normality made and discussed throughout the chapter can be tested on the standardized one-step ahead forecast errors (Durbin & Koopman, 2012, Chapter 7): $\tilde{\boldsymbol{v}}_t^k = \boldsymbol{B}_t^k \boldsymbol{v}_t^k$, for $t = d+1, \ldots, T$ with $\boldsymbol{B}_t^k$ such that $(\boldsymbol{F}_t^k)^{-1} = \boldsymbol{B}_t^{k'} \boldsymbol{B}_t^k$, where $\boldsymbol{F}_t^k$ is the covariance matrix of the prediction errors $\boldsymbol{v}_t^k$ estimated with the Kalman filter. The prediction errors for the labour force are defined as $\boldsymbol{v}_t^{k,y} = \boldsymbol{y}_t^k - \boldsymbol{Z}_t^y \hat{\boldsymbol{\alpha}}_{t|\Omega_{t-1}}^{k,y}$, for the claimant counts as $v_t^{k,CC} = x_t^{k,CC} - \boldsymbol{Z}^{CC} \hat{\boldsymbol{\alpha}}_{t|\Omega_{t-1}}^{k,CC}$, and for the Google Trends as $\boldsymbol{v}_t^{k,y} = \boldsymbol{x}_t^{k,GT} - \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{f}}_{t|\Omega_{t-1}}^k$, for $t = d+1, \ldots, T$ (the expressions for $\boldsymbol{Z}_t^y$ and $\boldsymbol{Z}^{CC}$ can be found in Appendix 2.A). We test the assumptions on the estimated CC & GT models when two factors of the Google Trends are included, and which employ, respectively, the monthly Google Trends, and the weekly

Google Trends with the additional iteration of the two-step estimator (as they yield the best results in terms of estimation and nowcast accuracy of the state variables of interest, when two factors of the Google Trends are included).

We test the null hypothesis of univariate normality for each of the prediction error, with the Shapiro and Wilk (1965) and Bowman and Shenton (1975) tests, as suggested, respectively, in Harvey (1989, Chapter 5) and Durbin and Koopman (2012, Chapter 2). The former test is based on the correlation between given observations and associated normal scores, whereas the latter test is based on the measures of skewness and kurtosis.

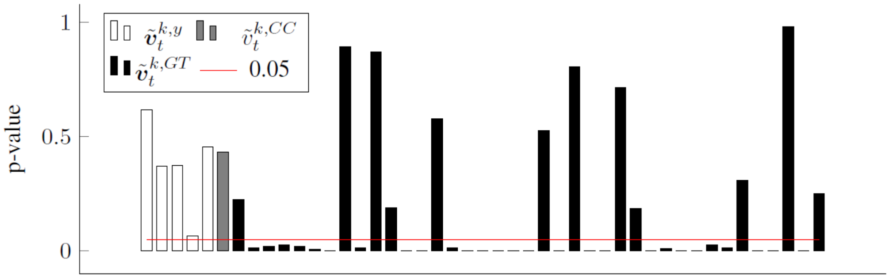


Figure 2.5: p-values from the Shapiro-Wilk test for individual normality, performed on each of the standardized prediction errors of the labour force, the claimant counts, and the Google Trends series ($\tilde{v}_t^k$). The standardized prediction errors are obtained from the CC & GT model which employs the monthly Google Trends and include two of their factors. The red line represents the confidence level of 0.05.

The p-values from the Shapiro-Wilk test are reported in Figures 2.5 and 2.6 for the two different model specifications discussed above, respectively. For both model specifications, there is no (strong) evidence against the normality assumptions for the error terms of the labour force and the claimant counts series, as their corresponding p-values are above the confidence level of 0.05. This result suggests that the model is correctly specified for these series. The test instead rejects the null hypothesis of normality for most of the idiosyncratic components of the Google Trends. The normality assumption seems therefore not appropriate for the latter series, but as discussed in Sections 2.3.1 and 2.5, and examined in the simulation study of Appendix 2.C, this

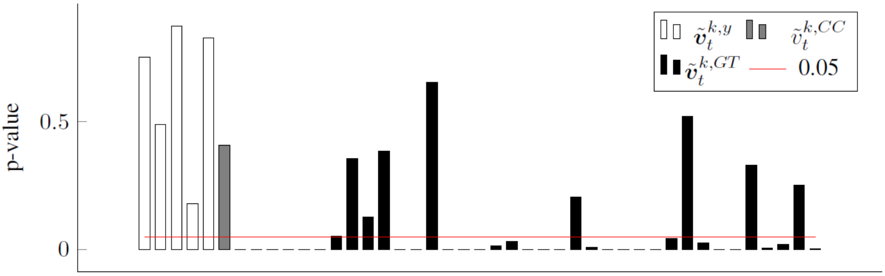


Figure 2.6: p-values from the Shapiro-Wilk test for individual normality, performed on each of the standardized prediction errors of the labour force, the claimant counts, and the Google Trends series ($\tilde{\boldsymbol{v}}_t^k$). The standardized prediction errors are obtained from the CC & GT model which employs the weekly Google Trends and include two of their factors, and which iterates the estimation of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. The red line represents the confidence level of 0.05.

type of misspecification does not affect the consistency of the estimators of the state variables and the hyperparameters, and does not seem to influence the performance of our method, nor the distribution of the LR test which allows us to perform inference on the correlation parameters[11]. The conclusions from the Bowman-Shenton test are the same and the corresponding p-values are reported in Figures 2.D.3 and 2.D.4.

[11] Notice that we do not control for multiple hypotheses testing in this case. If we did control for it, we would obtain fewer rejections of the null hypothesis of normality for the error terms of the Google Trends, but the conclusions for the error terms of the labour force and the claimant counts series would stay the same.

	Targeted GT, $n = 37, IC_1 = 1, IC_2 = 1, IC_3 = 2$							
	$r = 1$		$r = 2$					
					Weekly $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Psi}}$		Iterated $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Psi}}$	
	GT	CC & GT	GT	CC & GT	GT	CC & GT	GT	CC & GT
$\hat{\sigma}_{R,y}$	2020.195	2644.552	2590.937	3671.191	1995.064	2612.712	2238.557	2745.331
$\hat{\sigma}_{\omega,y}$	0.014	0.006	0.027	0.020	0.037	0.020	0.016	0.018
$\hat{\sigma}_{\lambda}$	3604.274	3738.299	3638.503	4281.357	3640.527	3568.508	3616.609	3635.421
$\hat{\sigma}_{\nu_1}$	1.146	1.151	1.142	1.181	1.161	1.147	1.155	1.148
$\hat{\sigma}_{\nu_2}$	1.295	1.286	1.292	1.376	1.294	1.294	1.278	1.312
$\hat{\sigma}_{\nu_3}$	1.203	1.171	1.208	1.211	1.167	1.204	1.207	1.199
$\hat{\sigma}_{\nu_4}$	1.253	1.225	1.248	1.358	1.247	1.274	1.252	1.267
$\hat{\sigma}_{\nu_5}$	1.240	1.179	1.234	1.227	1.244	1.225	1.231	1.243
$\hat{\delta}$	0.390	0.371	0.385	0.412	0.380	0.384	0.388	0.386
$\hat{\sigma}_{R,CC}$		3491.025		3635.234		3494.779		3508.248
$\hat{\sigma}_{\omega,CC}$		0.019		0.018		0.017		0.018
$\hat{\sigma}_{\varepsilon,CC}$		1342.202		1302.024		1280.971		1302.781
$\hat{\rho}_{CC}$		0.882		0.578		0.858		0.886
$\hat{\rho}_{1,GT}$	0.173	-0.054	0.441	-0.286	-0.101	-0.226	-0.245	0.275
$\hat{\rho}_{2,GT}$			0.539	-0.687	0.212	-0.030	0.371	0.015
$\widehat{\mathrm{MSE}}(\hat{L}^{k,y}_{t\vert\Omega_t})$	0.985	0.878	0.976	0.998	0.989	0.878	0.994	0.843
$\widehat{\mathrm{MSE}}(\hat{R}^{k,y}_{t\vert\Omega_t})$	0.936	0.872	0.904	0.996	0.912	0.836	0.961	0.799
$\widehat{\mathrm{MSE}}(\hat{\theta}^{k,y}_{t\vert\Omega_t})$	0.991	0.896	0.984	1.007	0.996	0.900	0.998	0.872
$\widehat{\mathrm{MSFE}}(\hat{L}^{k,y}_{t\vert\Omega_t^-})$	0.990	0.817	0.909	0.906	1.008	0.858	0.914	0.895
week 1	0.988	0.811	0.928	0.890	1.005	0.860	0.909	0.897
week 2	0.989	0.827	0.894	0.899	1.015	0.864	0.910	0.873
week 3	0.993	0.811	0.901	0.932	0.993	0.847	0.895	0.943
week 4	0.995	0.816	0.911	0.894	1.011	0.862	0.948	0.870
week 5	0.969	0.823	0.920	0.932	1.032	0.858	0.897	0.894
$\widehat{\mathrm{MSFE}}(\hat{R}^{k,y}_{t\vert\Omega_t^-})$	0.965	0.982	0.833	0.843	0.930	0.840	0.830	0.819
week 1	0.975	0.981	0.856	0.860	0.912	0.843	0.845	0.839
week 2	0.972	0.991	0.835	0.832	0.948	0.862	0.824	0.812
week 3	0.956	0.954	0.816	0.832	0.922	0.831	0.806	0.821
week 4	0.967	0.987	0.823	0.844	0.937	0.817	0.850	0.811
week 5	0.934	1.021	0.834	0.852	0.931	0.867	0.818	0.794
$\widehat{\mathrm{MSFE}}(\hat{\theta}^{k,y}_{t\vert\Omega_t^-})$	0.991	0.825	0.917	0.937	0.994	0.873	0.897	0.902
week 1	0.990	0.820	0.933	0.943	1.006	0.876	0.908	0.894
week 2	0.991	0.835	0.928	0.963	0.980	0.873	0.890	0.886
week 3	0.995	0.820	0.905	0.933	1.010	0.860	0.892	0.959
week 4	0.996	0.825	0.905	0.908	1.008	0.884	0.891	0.882
week 5	0.970	0.830	0.903	0.944	0.911	0.867	0.917	0.871
log-likelihood	-17954.398	-19621.000	-17767.456	-19438.409	-18651.434	-20318.457	-17745.335	-19413.916
	p-value from the LR test							
$H_0 : \rho_{CC} = 0$		0.001		0.000		0.001		0.000
$H_0 : \rho_{1,GT} = 0$	0.813	0.514	0.689	1.000	0.555	1.000	0.685	1.000
$H_0 : \rho_{2,GT} = 0$			0.133	0.070	0.604	1.000	0.221	1.000
$H_0 : \boldsymbol{\rho}_{GT} = \mathbf{0}$			0.247	0.062	0.759	1.000	0.429	1.000
$H_0 : \boldsymbol{\rho} = \mathbf{0}$		0.001		0.001		0.004		0.002

Table 2.4: Estimation and nowcast results for the labour force model with auxiliary series of claimant counts and weekly Google Trends about job-search and economic uncertainty (aggregated to the monthly frequency according to equation (2.3.4)). The number of Google Trends and the number of their factors included in the model are denoted with n and r, respectively. "Weekly $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Psi}}$" denotes that the latter estimates are obtained using the weekly Google Trends. "Iterated $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Psi}}$" means that the latter estimates are obtained from an additional iteration of the two-step estimator. "Targeted GT" indicates that the Google Trends have been targeted with the elastic net before including them in the model.

2.6 Conclusions

This chapter proposes a method to include a high-dimensional auxiliary series in a state space model in order to improve the estimation and nowcast of unobserved components. The method is based on a combination of PCA and Kalman filter estimation to reduce the dimensionality of the auxiliary series, originally proposed by Doz et al. (2011), while the auxiliary information is included in the state space model as in Harvey and Chung (2000). In this way we extend the state space model used by Statistics Netherlands to estimate the Dutch unemployment, which is based on monthly LFS data, by including the auxiliary series of claimant counts and Google Trends related to job-search and economic uncertainty. The strong explanatory power of the former series, in similar settings, has already been discovered in the literature (see Harvey and Chung (2000) and van den Brakel and Krieg (2016)). We explore to what extent a similar success can be obtained from online job-search and economic uncertainty behaviour. The advantage of Google Trends is that they are freely available at higher frequencies than the labour force survey and the claimant counts, and, contrary to the latter, they are not affected by publication delays. This feature can play a key role in the nowcasting of unemployment, as being the only real-time available information.

A Monte Carlo simulation study shows that in a smooth trend model our proposed method can improve the MSFE of the nowcasts of the trend's level and slope up to, respectively, around 25% and 75%. These results are robust to misspecifications regarding the distribution of the idiosyncratic components of the auxiliary series. Therefore, our method does have the potential to improve the nowcasts of unobserved components of interest.

In the empirical application of our method to Dutch unemployment estimation and nowcasting, we find that our considered Google Trends (when first targeted with the elastic net) do in general yield gains in the estimation and nowcast accuracy (respectively up to 40% and 25%) of the state variables of interest, with respect to the model which does not include any auxiliary series. This result stresses the advantage of using the high-dimensional auxiliary series of Google Trends, despite involving a more complex model to estimate, which is especially relevant for countries that do not have any data sources related to unemployment (such as the registry-sourced series of claimant counts), other than the labour force survey. We also find that, under certain model specifications, including both claimant counts and Google Trends outperforms the model which only includes the former auxiliary series. This result is explained by the fact that the two auxiliary series have a positive impact on the estimation/nowcast accuracy of different unobserved components which constitute the unemployment, thus yielding an overall improvement of the fit of the model. This also indicates that claimant counts and Google Trends do not bring redundant information about the Dutch unemployment.

The magnitude of the above-mentioned gains is, nonetheless, sensitive to the following aspects of the data and the model specification. First, in our empirical application we employ both monthly and weekly Google Trends. The latter need to be aggregated to the monthly frequency in order to be included in the model, but allow us to perform the nowcast on a weekly basis. We find that the former are less noisy and provide in general more accurate estimates/nowcasts of the state variables of interest. The explanatory power of the monthly Google Trends for Dutch unemployment is further corroborated by results from LR testing, which are in favour of their inclusion in the model. There is, instead, no strong and consistent evidence for this when the weekly Google Trends are employed.

Second, PCA involves the estimation of common factors that drive the Google Trends, and in our method we relate these factors to the unobserved components that constitute Dutch unemployment estimates. Information criteria suggest that the Google Trends are driven by either one or two common factors. We find that including two factors yields, in general, less gain in accuracy, compared to including one factor (due to the increased complexity of the model), but there is evidence that the second factor is related to the unemployment, and therefore it should be included in the model in order to exploit all the information that the Google Trends give about unemployment.

Finally, our estimation method is based on a two-step procedure. In the first step, the matrix of factor loadings and the covariance matrix of the idiosyncratic components of the Google Trends are estimated by PCA. In the second step, these matrices are replaced by their PCA estimates, in order to re-estimate the Google Trends' factors and the unobserved components of the labour force series, with the Kalman filter. Replacing these matrices by their estimates might affect the explanatory power of the Google Trends. We find that the explanatory power of the weekly Google Trends can be improved (in order to yield similar gains to the ones obtained with the monthly Google Trends), with an additional iteration of the two-step estimation procedure, which should provide more accurate estimates of the two matrices.

As already mentioned, we generally find estimation/nowcast accuracy gains from the inclusion of the Google Trends, when they are first "targeted", by selecting the ones that are relevant for Dutch unemployment, based on the elastic net penalized regression. If the targeting is not applied first, we do not find gains and significant relationships between the Google Trends and Dutch unemployment. Nonetheless, in this case the results do not deteriorate with respect to the model that does not include any auxiliary series, suggesting that our method is able to ignore the inclusion of irrelevant auxiliary series, in the estimation/nowcast of unobserved components of interest. This result is corroborated in our Monte Carlo simulation study. Hence, our proposed approach provides a framework to analyse the usefulness of "Big Data" sources, with little risk in case the series do not appear to be useful.

One limitation of the current chapter is that it does not allow for time-variation in the relation between the unobserved component of interest and the auxiliary series. For example, legislative changes may change the correlation between unemployment and administrative series such as claimant counts. Additionally, one can easily imagine the relevance of both specific search terms as well as internet search behaviour overall to change over time. While such time-variation may partly be addressed by considering shorter time periods, decreasing the already limited time dimension will have a strong detrimental effect on the quality of the estimators. Therefore, a more structural method is required that extends the current approach by building the potential for time variation into the estimation method directly, while retaining the possibility to use the full sample size. Such extensions are investigated in Chapters 3 and 4.

2.A State space representations

For the sake of simplicity, in this appendix material the subscript t (without the superscript k) indicates that the model is expressed at the low (monthly) frequency.

2.A.1 Labour force model with univariate auxiliary series

Throughout this section it is assumed that the univariate auxiliary series are the claimant counts, therefore $x_t = x_t^{CC}$.

The observation equation is:

$$\underset{6\times1}{\begin{pmatrix} \boldsymbol{y}_t \\ x_t \end{pmatrix}} = \boldsymbol{Z}_t \begin{pmatrix} \boldsymbol{\alpha}_t^y \\ \boldsymbol{\alpha}_t^x \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \varepsilon_t^x \end{pmatrix} = \begin{bmatrix} \boldsymbol{Z}_t^y & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}^x \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_t^y \\ \boldsymbol{\alpha}_t^x \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \varepsilon_t^x \end{pmatrix},$$

$$\begin{pmatrix} \mathbf{0} \\ \varepsilon_t^x \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{H}),$$

$$\underset{6\times6}{\boldsymbol{H}} = \operatorname{diag}\left(\mathbf{0}', \sigma_{\varepsilon,x}^2\right).$$

The state variables for $\boldsymbol{y}_t$ (i.e., the level, the slope, the seasonality, the RGB and the survey errors) are:

$$\underset{30\times1}{\boldsymbol{\alpha}_t^y} = \big(L_t^y \quad R_t^y \quad S_{1,t}^y \quad S_{1,t}^{*y} \quad S_{2,t}^y \quad S_{2,t}^{*y} \quad S_{3,t}^y \quad S_{3,t}^{*y} \quad S_{4,t}^y \quad S_{4,t}^{*y} \\ S_{5,t}^y \quad S_{5,t}^{*y} \quad S_{6,t}^y \quad \lambda_{2,t} \quad \lambda_{3,t} \quad \lambda_{4,t} \quad \lambda_{5,t} \quad \boldsymbol{\alpha}'_{E,t}\big)'$$

$$\underset{13\times1}{\boldsymbol{\alpha}_{E,t}} = \big(\tilde{e}_{1,t} \quad \tilde{e}_{2,t} \quad \tilde{e}_{3,t} \quad \tilde{e}_{4,t} \quad \tilde{e}_{5,t} \quad \tilde{e}_{1,t-2} \quad \tilde{e}_{2,t-2} \quad \tilde{e}_{3,t-2} \quad \tilde{e}_{4,t-2} \\ \tilde{e}_{1,t-1} \quad \tilde{e}_{2,t-1} \quad \tilde{e}_{3,t-1} \quad \tilde{e}_{4,t-1}\big)',$$

where E refers to the structure of the autocorrelated sampling errors that are modelled as state variables.

The state variables for x_t (i.e., the level, the slope and the seasonality) are:

$$\underset{13\times 1}{\boldsymbol{\alpha}_t^x} = \big(L_t^x \quad R_t^x \quad S_{1,t}^x \quad S_{1,t}^{*x} \quad S_{2,t}^x \quad S_{2,t}^{*x} \quad S_{3,t}^x \quad S_{3,t}^{*x} \quad S_{4,t}^x \quad S_{4,t}^{*x} \\ S_{5,t}^x \quad S_{5,t}^{*x} \quad S_{6,t}^x\big)'.$$

$$\underset{5\times 30}{\boldsymbol{Z}_t^y} = \left[\begin{array}{ccccccccccccccccc c} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & \boldsymbol{Z}_{E,t}^y \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & \end{array}\right],$$

$$\underset{5\times 13}{\boldsymbol{Z}_{E,t}^y} = \begin{bmatrix} c_{1,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c_{2,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & c_{3,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{4,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{5,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\underset{1\times 13}{\boldsymbol{Z}^x} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

The transition equation takes the form:

$$\underset{43\times 1}{\begin{pmatrix} \boldsymbol{\alpha}_t^y \\ \boldsymbol{\alpha}_t^x \end{pmatrix}} = \boldsymbol{T}\begin{pmatrix} \boldsymbol{\alpha}_{t-1}^y \\ \boldsymbol{\alpha}_{t-1}^x \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^y \\ \boldsymbol{\eta}_t^x \end{pmatrix} = \begin{bmatrix} \boldsymbol{T}^y & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T}^x \end{bmatrix}\begin{pmatrix} \boldsymbol{\alpha}_{t-1}^y \\ \boldsymbol{\alpha}_{t-1}^x \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^y \\ \boldsymbol{\eta}_t^x \end{pmatrix}.$$

The transition matrix for $\boldsymbol{y}_t$ is:

$$\underset{30\times 30}{\boldsymbol{T}^y} = \text{blockdiag}(\boldsymbol{T}_\mu^y, \boldsymbol{T}_\omega^y, \boldsymbol{T}_\lambda^y, \boldsymbol{T}_E^y).$$

The transition matrix for the level and slope components is:

$$\underset{2\times 2}{\boldsymbol{T}_\mu^y} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

The transition matrix for the seasonal component is:

$$\underset{11\times 11}{\boldsymbol{T}^y_\omega} = \text{blockdiag}(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{C}_4, \boldsymbol{C}_5, -1),$$

$$\boldsymbol{C}_l = \begin{bmatrix} \cos(h_l) & \sin(h_l) \\ -\sin(h_l) & \cos(h_l) \end{bmatrix}, \quad h_l = \pi l/6, \quad l = 1, ..., 5.$$

The transition matrix for the RGB component is:

$$\underset{4\times 4}{\boldsymbol{T}^y_\lambda} = \boldsymbol{I}_4.$$

The transition matrix for the autocorrelated survey errors is:

$$\underset{13\times 13}{\boldsymbol{T}^y_E} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The transition matrix for x_t, $\underset{13\times 13}{\boldsymbol{T}^x} = \text{blockdiag}\,(\boldsymbol{T}^y_\mu, \boldsymbol{T}^y_\omega)$, is the same as $\boldsymbol{T}^y$ without the transition matrices for the RGB component and for the survey errors.

The vector of innovations is defined as follows:

$$\begin{aligned}
\underset{30\times1}{\boldsymbol{\eta}_t^y} &= \big(\eta^y_{L,t} \quad \eta^y_{R,t} \quad \eta^y_{\omega,1,t} \quad \eta^{*y}_{\omega,1,t} \quad \eta^y_{\omega,2,t} \quad \eta^{*y}_{\omega,2,t} \quad \eta^y_{\omega,3,t} \quad \eta^{*y}_{\omega,3,t} \quad \eta^y_{\omega,4,t} \\
&\qquad \eta^{*y}_{\omega,4,t} \quad \eta^y_{\omega,5,t} \quad \eta^{*y}_{\omega,5,t} \quad \eta^y_{\omega,6,t} \quad \eta_{\lambda,2,t} \quad \eta_{\lambda,3,t} \quad \eta_{\lambda,4,t} \quad \eta_{\lambda,5,t} \quad \boldsymbol{\eta}'^y_{E,t}\big)', \\
\underset{13\times1}{\boldsymbol{\eta}_{E,t}^y} &= \big(\ \nu_{1,t} \quad \nu_{2,t} \quad \nu_{3,t} \quad \nu_{4,t} \quad \nu_{5,t} \quad \mathbf{0}'\ \big)', \\
\underset{13\times1}{\boldsymbol{\eta}_t^x} &= \big(\eta^x_{L,t} \quad \eta^x_{R,t} \quad \eta^x_{\omega,1,t} \quad \eta^{*x}_{\omega,1,t} \quad \eta^x_{\omega,2,t} \quad \eta^{*x}_{\omega,2,t} \quad \eta^x_{\omega,3,t} \quad \eta^{*x}_{\omega,3,t} \quad \eta^x_{\omega,4,t} \\
&\qquad \eta^{*x}_{\omega,4,t} \quad \eta^x_{\omega,5,t} \quad \eta^{*x}_{\omega,5,t} \quad \eta^x_{\omega,6,t}\big)', \\
\underset{43\times1}{\boldsymbol{\eta}_t} &= \big(\ \boldsymbol{\eta}_t'^y \quad \boldsymbol{\eta}_t'^x\ \big)' \sim N\left(\mathbf{0}, \boldsymbol{Q}\right),
\end{aligned}$$

$$\underset{43\times43}{\boldsymbol{Q}} =$$

$$= \begin{bmatrix}
\sigma^2_{L,y} & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & 0 & \mathbf{0}' \\
0 & \sigma^2_{R,y} & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & \rho\sigma_{R,y}\sigma_{R,x} & \mathbf{0}' \\
\mathbf{0} & \mathbf{0} & \boldsymbol{Q}^y_\omega & \underset{11\times4}{\mathbf{0}} & \underset{11\times5}{\mathbf{0}} & \underset{11\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{11\times11}{\mathbf{0}} \\
\mathbf{0} & \mathbf{0} & \underset{4\times11}{\mathbf{0}} & \boldsymbol{Q}^y_\lambda & \underset{4\times5}{\mathbf{0}} & \underset{4\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{4\times11}{\mathbf{0}} \\
\mathbf{0} & \mathbf{0} & \underset{5\times11}{\mathbf{0}} & \underset{5\times4}{\mathbf{0}} & \boldsymbol{Q}^y_\nu & \underset{5\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{5\times11}{\mathbf{0}} \\
\mathbf{0} & \mathbf{0} & \underset{8\times11}{\mathbf{0}} & \underset{8\times4}{\mathbf{0}} & \underset{8\times5}{\mathbf{0}} & \underset{8\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{8\times11}{\mathbf{0}} \\
0 & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \sigma^2_{L,x} & 0 & \mathbf{0}' \\
0 & \rho\sigma_{R,y}\sigma_{R,x} & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & \sigma^2_{R,x} & \mathbf{0}' \\
\mathbf{0} & \mathbf{0} & \underset{11\times11}{\mathbf{0}} & \underset{11\times4}{\mathbf{0}} & \underset{11\times5}{\mathbf{0}} & \underset{11\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \boldsymbol{Q}^x_\omega
\end{bmatrix},$$

where $\sigma^2_{L,y} = \sigma^2_{L,x} = 0$ in the Dutch labour force model, $\underset{11\times11}{\boldsymbol{Q}^z_\omega} = \sigma^2_{\omega,z}\boldsymbol{I}_{11}$, for $z = x, y$, $\underset{4\times4}{\boldsymbol{Q}^y_\lambda} = \sigma^2_\lambda \boldsymbol{I}_4$ and $\underset{5\times5}{\boldsymbol{Q}^y_\nu} = \text{diag}\left(\sigma^2_{\nu_1}, \sigma^2_{\nu_2}, \sigma^2_{\nu_3}, \sigma^2_{\nu_4}, \sigma^2_{\nu_5}\right)$.

2.A.2 Labour force model with high-dimensional auxiliary series

Throughout this section it is assumed that the high-dimensional auxiliary series are the Google Trends, therefore $\boldsymbol{x}_t = \boldsymbol{x}_t^{GT}$. n is the number of Google Trends. It is assumed only $r = 1$ factor for the Google Trends.

The observation equation is:

$$\underset{(5+n)\times 1}{\begin{pmatrix} \boldsymbol{y}_t \\ \boldsymbol{x}_t \end{pmatrix}} = \underset{(5+n)\times 31}{\boldsymbol{Z}_t} \begin{pmatrix} \boldsymbol{\alpha}_t^y \\ \alpha_t^x \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\varepsilon}_t \end{pmatrix} =$$

$$= \begin{bmatrix} \boldsymbol{Z}_t^y & \mathbf{0} \\ \underset{n\times 31}{\mathbf{0}} & \underset{n\times 1}{\hat{\Lambda}} \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_t^y \\ f_t \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\varepsilon}_t \end{pmatrix}, \quad \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\varepsilon}_t \end{pmatrix} \sim N\left(\mathbf{0}, \hat{\boldsymbol{H}}\right),$$

$$\underset{(5+n)\times(5+n)}{\hat{\boldsymbol{H}}} = \operatorname{diag}\left(\mathbf{0}', \hat{\psi}_{11}, \ldots, \hat{\psi}_{nn}\right).$$

$\boldsymbol{Z}_t^y$ is the same as in Appendix 2.A.1.

The transition equation takes the form:

$$\underset{31\times 1}{\begin{pmatrix} \boldsymbol{\alpha}_t^y \\ f_t \end{pmatrix}} = \underset{31\times 31}{\boldsymbol{T}} \begin{pmatrix} \boldsymbol{\alpha}_{t-1}^y \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^y \\ \eta_t^x \end{pmatrix} =$$

$$= \begin{bmatrix} \boldsymbol{T}^y & \mathbf{0} \\ \mathbf{0} & \boldsymbol{T}^x \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_{t-1}^y \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^y \\ u_t \end{pmatrix}.$$

$\boldsymbol{T}^y$ is the same as in Appendix 2.A.1, and $\boldsymbol{T}^x = 1$.

The vector of innovations is:

$$\underset{31\times 1}{\boldsymbol{\eta}_t} = \left(\begin{array}{cc} \boldsymbol{\eta}_t^{\prime y} & u_t \end{array} \right)' \sim N\left(\mathbf{0}, \boldsymbol{Q}\right),$$

$$\underset{31\times 31}{\boldsymbol{Q}} = \left[\begin{array}{ccccccc} \sigma^2_{L,y} & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 \\ 0 & \sigma^2_{R,y} & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \rho\sigma_{R,y}\sigma_u \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Q}^y_\omega & \underset{11\times 4}{\mathbf{0}} & \underset{11\times 5}{\mathbf{0}} & \underset{11\times 8}{\mathbf{0}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{4\times 11}{\mathbf{0}} & \boldsymbol{Q}^y_\lambda & \underset{4\times 5}{\mathbf{0}} & \underset{4\times 8}{\mathbf{0}} & \mathbf{0} \\ 0 & 0 & \underset{5\times 11}{\mathbf{0}} & \underset{5\times 4}{\mathbf{0}} & \boldsymbol{Q}^y_\nu & \underset{5\times 8}{\mathbf{0}} & 0 \\ \mathbf{0} & \mathbf{0} & \underset{8\times 11}{\mathbf{0}} & \underset{8\times 4}{\mathbf{0}} & \underset{8\times 5}{\mathbf{0}} & \underset{8\times 8}{\mathbf{0}} & \mathbf{0} \\ 0 & \rho\sigma_{R,y}\sigma_u & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \sigma^2_u \end{array} \right],$$

where $\boldsymbol{\eta}_t^y$ and the first (30×30) diagonal elements of $\boldsymbol{Q}$ are the same as in Appendix 2.A.1.

Extension of the model to incorporate the lags of f_t

Consider a regression of $\eta^y_{R,t}$ on the past values of u_t:

$$\left(\begin{array}{c} R^y_t \\ f_t \end{array} \right) = \left(\begin{array}{c} R^y_{t-1} \\ f_{t-1} \end{array} \right) + \left(\begin{array}{c} \eta^y_{R,t} \\ u_t \end{array} \right), \quad u_t \sim N\left(0, \sigma^2_u\right),$$

$$\eta^y_{R,t} = \sum_{j=1}^{q} \kappa_j u_{t-j} + w_t =$$

$$= \kappa_1 f_{t-1} + \sum_{j=2}^{q} \left(\kappa_j - \kappa_{j-1}\right) f_{t-j} - \kappa_q f_{t-q-1} + w_t,$$

$$w_t \sim N\left(0, \sigma^2_w\right).$$

$$
\begin{pmatrix} L_t^y \\ R_t^y \\ f_t \\ f_{t-1} \\ f_{t-2} \\ \vdots \\ f_{t-q} \end{pmatrix} =
$$

$$
= \begin{bmatrix} 1 & 1 & 0 & 0 & \mathbf{0}' & 0 & 0 \\ 0 & 1 & \kappa_1 & (\kappa_2 - \kappa_1) & \dots & (\kappa_q - \kappa_{q-1}) & -\kappa_q \\ 0 & 0 & 1 & 0 & \mathbf{0}' & 0 & 0 \\ 0 & 0 & 1 & 0 & \mathbf{0}' & 0 & 0 \\ 0 & 0 & 0 & 1 & \mathbf{0}' & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\ 0 & 0 & 0 & 0 & \mathbf{0}' & 1 & 0 \end{bmatrix} \begin{pmatrix} L_{t-1}^y \\ R_{t-1}^y \\ f_{t-1} \\ f_{t-2} \\ f_{t-3} \\ \vdots \\ f_{t-q-1} \end{pmatrix} +
$$

$$
+ \begin{pmatrix} 0 \\ w_t \\ u_t \\ 0 \\ 0 \\ 0 \\ \mathbf{0} \\ 0 \end{pmatrix}, \quad \begin{pmatrix} w_t \\ u_t \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_w^2 & \rho\sigma_w\sigma_u \\ \rho\sigma_w\sigma_u & \sigma_u^2 \end{bmatrix}\right).
$$

In the measurement equation $\boldsymbol{Z}^x = \begin{bmatrix} \underset{n\times 1}{\hat{\Lambda}} & \underset{n\times q}{\mathbf{0}} \end{bmatrix}$.

Extension of the model to incorporate the seasonality/cycle in f_t with a (seasonal) ARIMA model

Assume an ARIMA$(3,1,1)$ process for f_t:

$$\begin{aligned} f_t = f_{t-1} + \phi_1(f_{t-1} - f_{t-2}) + \phi_2(f_{t-2} - f_{t-3}) + \phi_3(f_{t-3} - f_{t-4}) + \\ + u_t + \gamma u_{t-1}, \quad u_t \sim N(0,1). \end{aligned}$$

The state space representation of the above model is based on Durbin and Koopman (2012, Chapter 3) and illustrated below. Let $\boldsymbol{f}_t$ be the state vector

$$\boldsymbol{f}_t = \begin{pmatrix} f_{t-1} \\ f_t - f_{t-1} \\ \phi_2(f_{t-1} - f_{t-2}) \\ \phi_3(f_{t-2} - f_{t-3}) + \gamma u_t \end{pmatrix}.$$

The transition equation for $\boldsymbol{f}_t$ takes the form:

$$\boldsymbol{f}_t = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & \phi_1 & 1 & 1 \\ 0 & \phi_2 & 0 & 0 \\ 0 & 0 & \frac{\phi_3}{\phi_2} & 0 \end{bmatrix} \boldsymbol{f}_{t-1} + \begin{pmatrix} 0 \\ 1 \\ 0 \\ \gamma \end{pmatrix} u_t.$$

Consequently, the observation equation becomes:

$$\boldsymbol{x}_t = \hat{\boldsymbol{\Lambda}} \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix} + \boldsymbol{\varepsilon}_t.$$

Note that the transition equation of the full state space model is now expressed in the form:

$$\boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{R}\boldsymbol{\eta}_t,$$

where

$$\underset{\dim(\boldsymbol{\alpha}_t)\times\dim(\boldsymbol{\alpha}_t)}{\boldsymbol{R}} = \begin{bmatrix} \boldsymbol{I}_{\dim(\boldsymbol{\alpha}_t)-4} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & 0 & 0 & 0 & 0 \\ \mathbf{0}' & 0 & 1 & 0 & 0 \\ \mathbf{0}' & 0 & 0 & 0 & 0 \\ \mathbf{0}' & 0 & 0 & 0 & \gamma \end{bmatrix}.$$

We here allow u_t to be correlated with $\eta^y_{R,t}$.

I(1) idiosyncratic components

Consider the following toy example to have a clearer understanding of the estimation procedure when some of the idiosyncratic components are $I(1)$.

$$\boldsymbol{x}_t = \Lambda f_t + \boldsymbol{\varepsilon}_t.$$

Suppose that $\boldsymbol{x}_t$ and $\boldsymbol{\varepsilon}_t$ are 5-dimensional vectors ($n = 5$), and f_t is univariate. Suppose that $\varepsilon_{1,t}$ and $\varepsilon_{3,t}$ are $I(1)$, whereas $\varepsilon_{2,t}$, $\varepsilon_{4,t}$ and $\varepsilon_{5,t}$ are $I(0)$. Then the observation equation for $\boldsymbol{x}_t$ becomes:

$$\begin{pmatrix} x_{1,t} \\ x_{2,t} \\ x_{3,t} \\ x_{4,t} \\ x_{5,t} \end{pmatrix} = \begin{bmatrix} \Lambda_1 & 1 & 0 \\ \Lambda_2 & 0 & 0 \\ \Lambda_3 & 0 & 1 \\ \Lambda_4 & 0 & 0 \\ \Lambda_5 & 0 & 0 \end{bmatrix} \begin{pmatrix} f_t \\ \varepsilon_{1,t} \\ \varepsilon_{3,t} \end{pmatrix} + \begin{pmatrix} 0 \\ \varepsilon_{2,t} \\ 0 \\ \varepsilon_{4,t} \\ \varepsilon_{5,t} \end{pmatrix},$$

where f_t, $\varepsilon_{1,t}$ and $\varepsilon_{3,t}$ are state variables with transition equation

$$\begin{pmatrix} f_t \\ \varepsilon_{1,t} \\ \varepsilon_{3,t} \end{pmatrix} = \boldsymbol{I}_3 \begin{pmatrix} f_{t-1} \\ \varepsilon_{1,t-1} \\ \varepsilon_{3,t-1} \end{pmatrix} + \begin{pmatrix} u_t \\ \xi_{1,t} \\ \xi_{3,t} \end{pmatrix}.$$

$$\begin{aligned} \boldsymbol{\Psi} &= \operatorname{cov}\begin{pmatrix} \xi_{1,t} & \varepsilon_{2,t} & \xi_{3,t} & \varepsilon_{4,t} & \varepsilon_{5,t} \end{pmatrix}' = \\ &= \operatorname{cov}\begin{pmatrix} \Delta\varepsilon_{1,t} & \varepsilon_{2,t} & \Delta\varepsilon_{3,t} & \varepsilon_{4,t} & \varepsilon_{5,t} \end{pmatrix}'. \end{aligned}$$

The covariance matrix between the innovation terms in the observation equation is

$$\operatorname{cov}\begin{pmatrix} 0 & \varepsilon_{2,t} & 0 & \varepsilon_{4,t} & \varepsilon_{5,t} \end{pmatrix}' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \psi_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \psi_{44} & 0 \\ 0 & 0 & 0 & 0 & \psi_{55} \end{pmatrix},$$

and ends up in the $\boldsymbol{H}$ matrix defined in Appendices 2.A.2 or 2.A.3. On the contrary, the covariance matrix between the innovations of the state variables is

$$\operatorname{cov}\begin{pmatrix} u_t & \xi_{1,t} & \xi_{3,t} \end{pmatrix}' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \psi_{11} & 0 \\ 0 & 0 & \psi_{33} \end{pmatrix},$$

and ends up in the $\boldsymbol{Q}$ matrix defined in Appendices 2.A.2 or 2.A.3.

2.A.3 Labour force model with univariate and high-dimensional auxiliary series

Throughout this section both the claimant counts and the Google Trends are included in the model as auxiliary series.

The observation equation is:

$$\underset{(6+n)\times 1}{\begin{pmatrix} \boldsymbol{y}_t \\ x_t^{CC} \\ \boldsymbol{x}_t^{GT} \end{pmatrix}} = \underset{(6+n)\times 44}{\boldsymbol{Z}_t} \begin{pmatrix} \boldsymbol{\alpha}_t^{y} \\ \boldsymbol{\alpha}_t^{CC} \\ \alpha_t^{GT} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \varepsilon_t^{CC} \\ \boldsymbol{\varepsilon}_t \end{pmatrix} =$$

$$= \begin{bmatrix} \boldsymbol{Z}_t^{y} & \underset{5\times 13}{\boldsymbol{0}} & \boldsymbol{0} \\ \boldsymbol{0}' & \underset{1\times 13}{\boldsymbol{Z}^{CC}} & 0 \\ \underset{n\times 30}{\boldsymbol{0}} & \underset{n\times 13}{\boldsymbol{0}} & \underset{n\times 1}{\hat{\Lambda}} \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_t^{y} \\ \boldsymbol{\alpha}_t^{CC} \\ f_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \varepsilon_t^{CC} \\ \boldsymbol{\varepsilon}_t^{GT} \end{pmatrix},$$

$$\begin{pmatrix} \boldsymbol{0} \\ \varepsilon_t^{CC} \\ \boldsymbol{\varepsilon}_t^{GT} \end{pmatrix} \sim N\left(\boldsymbol{0}, \boldsymbol{H}\right), \qquad \underset{(6+n)\times(6+n)}{\boldsymbol{H}} = \operatorname{diag}\left(\boldsymbol{0}', \sigma^2_{\varepsilon,x}, \hat{\psi}_{11}, \ldots, \hat{\psi}_{nn}\right).$$

$\boldsymbol{Z}_t^{y}$ is the same as in Appendix 2.A.1, and $\boldsymbol{Z}^{CC}$ is the same as $\boldsymbol{Z}^{x}$ in Appendix 2.A.1.

The transition equation takes the form:

$$\underset{44\times 1}{\begin{pmatrix} \boldsymbol{\alpha}_t^{y} \\ \boldsymbol{\alpha}_t^{CC} \\ f_t \end{pmatrix}} = \underset{44\times 44}{\boldsymbol{T}} \begin{pmatrix} \boldsymbol{\alpha}_{t-1}^{y} \\ \boldsymbol{\alpha}_{t-1}^{CC} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^{y} \\ \boldsymbol{\eta}_t^{CC} \\ \eta_t^{GT} \end{pmatrix} =$$

$$= \begin{bmatrix} \boldsymbol{T}^{y} & \underset{30\times 13}{\boldsymbol{0}} & \boldsymbol{0} \\ \underset{13\times 30}{\boldsymbol{0}} & \boldsymbol{T}^{CC} & \boldsymbol{0} \\ \boldsymbol{0}' & \boldsymbol{0}' & 1 \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_{t-1}^{y} \\ \boldsymbol{\alpha}_{t-1}^{CC} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^{y} \\ \boldsymbol{\eta}_t^{CC} \\ u_t \end{pmatrix}.$$

$\boldsymbol{T}^{y}$ is the same as in Appendix 2.A.1, and $\boldsymbol{T}^{CC}$ and $\boldsymbol{\alpha}_t^{CC}$ are, respectively, the same as $\boldsymbol{T}^{x}$ and $\boldsymbol{\alpha}_t^{x}$ in Appendix 2.A.1.

The vector of innovations is:

$$\underset{44\times 1}{\boldsymbol{\eta}_t} = \begin{pmatrix} \boldsymbol{\eta}_t^{\prime y} & \boldsymbol{\eta}_t^{\prime CC} & u_t \end{pmatrix}' \sim N\left(\mathbf{0}, \boldsymbol{Q}\right),$$

$$\underset{44\times 44}{\boldsymbol{Q}} =$$

$$= \begin{bmatrix} \sigma_{L,y}^2 & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & 0 & \mathbf{0}' & 0 \\ 0 & \sigma_{R,y}^2 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & \rho_{CC}\sigma_{R,y}\sigma_{R,CC} & \mathbf{0}' & \rho_{GT}\sigma_{R,y}\sigma_u \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Q}_\omega^y & \underset{11\times 4}{\mathbf{0}} & \underset{11\times 5}{\mathbf{0}} & \underset{11\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{11\times 11}{\mathbf{0}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{4\times 11}{\mathbf{0}} & \boldsymbol{Q}_\lambda^y & \underset{4\times 5}{\mathbf{0}} & \underset{4\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{4\times 11}{\mathbf{0}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{5\times 11}{\mathbf{0}} & \underset{5\times 4}{\mathbf{0}} & \boldsymbol{Q}_\nu^y & \underset{5\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{5\times 11}{\mathbf{0}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{8\times 11}{\mathbf{0}} & \underset{8\times 4}{\mathbf{0}} & \underset{8\times 5}{\mathbf{0}} & \underset{8\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{8\times 11}{\mathbf{0}} & \mathbf{0} \\ 0 & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \sigma_{L,CC}^2 & 0 & \mathbf{0}' & 0 \\ 0 & \rho_{CC}\sigma_{R,y}\sigma_{R,CC} & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & \sigma_{R,CC}^2 & \mathbf{0}' & 0 \\ \mathbf{0} & \mathbf{0} & \underset{11\times 11}{\mathbf{0}} & \underset{11\times 4}{\mathbf{0}} & \underset{11\times 5}{\mathbf{0}} & \underset{11\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \boldsymbol{Q}_\omega^{CC} & \mathbf{0} \\ 0 & \rho_{GT}\sigma_{R,y}\sigma_u & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & 0 & \mathbf{0}' & \sigma_u^2 \end{bmatrix},$$

where $\boldsymbol{\eta}_t^y$ is the same as in Appendix 2.A.1. $\boldsymbol{\eta}_t^{CC}$ and $\sigma_{R,CC}$ are respectively the same as $\boldsymbol{\eta}_t^x$ and $\sigma_{R,x}$ in Appendix 2.A.1. The first (43×43) elements of $\boldsymbol{Q}$ are the same as in Appendix 2.A.1, whereas the last row and column are the same as in Appendix 2.A.2.

2.B List of Google Trends

Search term	Translation/explanation	Search term	Translation/explanation
aanpassing	adjustment	job interview	
aanvragen uitkering	to apply for benefit	job vacancy	
adecco	Adecco is an employment agency	jobbird	Jobbird is a website to look for a job
advertentie plaatsen	to place an advertisement	jobbird vacatures	Jobbird vacancies
adverteren	to announce	jobnet	Jobnet is a website to look for a job
arbeidsbureau	employment office	jobs	
automatische incasso	automatic collection of money	jobtrack	
baan	job	juridische vacatures	legal vanacies
baan zoeken	job search	kantonrechter	cantonal judge
banen	jobs	kantonrechtersformule	cantonal court formula (to treat e.g. severance payments)
bedrijven failliet	businesses bankrupt	maatschappelijk werk	social work
belegger.nl	website about investments' information	manpower	Manpower is an employment agency
bezuinigen	to economize	maximum dagloon	maximum daily wage
bijscholen	retraining	mijn uwv	my uwv
bijstand	assistance	modernisering	modernization
bijstandsuitkering	social assistance benefit	monsterboard	Monsterboard is a website to look for a job
collectief ontslag	collective dismissal	monsterboard vacatures	Monsterboard vacancies
creyfs	Creyfs is an empolyment agency	motivatiebrief	motivation letter
curriculum vitae		motivatiebrief schrijven	to write a motivation letter
curriculum vitae template		motivatiebrief voorbeeld	example of motivation letter
curriculum vitae voorbeeld	curriculum vitae example	motivation letter	
cv		nationale vacaturebank	national job bank
cv maken	to make a cv	olympia uitzendbureau	Olympia employment agency
cv maken voorbeeld	to make a cv example	omscholen	retrain
dagloon	daily wage	ondernemingsplan voorbeeld	business plan example
duur ww	duration of the unemployment benefit	ontslag	dismissal
economische crisis	economic crisis	ontslagaanvraag	dismissal application
failliet	bankrupt	ontslagprocedure	dismissal procedure
faillisementen	bankruptcies	ontslagvergoeding	severance pay
fulltime baan	full-time job	ontslagvergunning	dismissal permit
functieomschrijving	job description	open sollicitatiebrief	open application letter
geen werk	no work	partijhandel	stock trading
hoofdbedrijfschap ambachten	main business crafts	productiemedewerker	production employee
hoogte ww	level of unemployment benefit	promotiewerk	promotional work
huizenmarkt zeepbel	housing market bubble	randstad	Randstad is an employment agency
ict vacatures	IT vacancies	randstad jobs	
ik zoek werk	I am looking for a job	randstad uitzendbureau	Randstad employment agency
indeed	Indeed is a website to look for a job	randstad vacatures	Randstad vacancies
indeed jobs		receptioniste	receptionist
indeed uitzendbureau	Indeed employment agency	recht op ww	right to unemployment
indeed vacatures	Indeed vacancies	reorganisatie	reorganization
ing direct	website of the ING bank	restructuring	
interim	Interim is an employment agency	resume	
job		resumé	
job bird	Jobbird is a website to look for a job	resume example	
job description		resume template	
		salarisadministrateur	payroll administrator

Table 2.B.1: List of Google search terms and their translations/explanations (part 1).

Search term	Translation/explanation	Search term	Translation/explanation
schoonmaakwerk	cleaning work	vacatures limburg	jobs in Limburg (Dutch province)
schuldsanering	debt restructuring	vacatures noord brabant	jobs in North Brabant (Dutch province)
sociaal plan	social plan	vacatures zorg	vacancies care
sollicitatie	job application	vakantiebaan	vacation job
sollicitatiebrief	cover letter	vakantiewerk	holidayjob
sollicitatiebrief schrijven	to write a cover letter	verkoopmedewerker	sales employee
sollicitatiebrief voorbeeld	cover letter example	voorbeeld cv	example cv
sollicitatiegesprek	job interview	voorbeeld motivatiebrief	example of motivation letter
sollicitaties	job applications	vrijwilligerswerk	volunteer work
solliciteren	to apply	vrijwilligerswerk buitenland	volunteering abroad
solliciteren bij	apply at	werk gezocht	job search
start people	Start People is an employment agency	werk in	work in
start uitzendbureau	Start employment agency	werk nl	work NL
tempo team	Tempo Team is an employment agency	werk.nl	website for job placement
tempo-team	Tempo Team is an employment agency	werk rotterdam	work Rotterdam
tempo team uitzendbureau	Tempo Team employment agency	werk utrecht	work Utrecht
tempo team vacatures	Tempo Team vacancies	werk vacature	job vacancy
tempoteam	Tempo Team is an employment agency	werk vacatures	job vacancies
tence	Tence is an employment agency	werk vinden	to find a job
tijdelijk werk	temporary job	werk zoeken	to look for a job
uitkering	payment	werkbedrijf	operating company
uitkering aanvragen	to claim benefits	werkeloos	unemployed
uitzendbureau	employment agency	werken bij	to work at
uitzendbureau amsterdam	employment agency Amsterdam	werken in	to work in
uitzendbureau den haag	employment agency The Hague	werking	working
uitzendbureaus	employment agencies	werkloos	unemployed
uwv	Employee Insurance Agency	werkloosheid	unemployment
uwv uitkering	Employee Insurance Agency payment	werkloosheidsuitkering	unemployment benefits
uwv vacatures	Employee Insurance Agency vacancies	werkloosheidswet	unemployment law
uwv werkbedrijf	Employee Insurance Agency operating company	werkloze	unemployed person
uwv ww	Employee Insurance Agency unemployment benefits	werklozen	unemployed people
vacature	job offer	werkzoekende	job seeker
vacature amsterdam	job offer Amsterdam	wet op de ondernemingsraden	Works Councils Act
vacature eindhoven	job offer Eindhoven	wholesale	
vacature secretaresse	vacancy secretary	Ww	unemployment benefits
vacaturebank	job bank	ww	unemployment benefits
vacatures	job offers	ww aanvragen	to request unemployment benefits
vacatures beveiliging	job security	ww uitkering	unemployment benefit payments
vacatures bouw	job construction	ww-uitkering	unemployment benefit payments
vacatures brabant	jobs in Brabant (Dutch province)	ww uitkering aanvragen	claim benefits
vacatures communicatie	vacancies communication	ww uitkering aanvragen uwv	claim benefits Employee Insurance Agency
vacatures flevoland	jobs in Flevoland (Dutch province)	www.asnbank.nl	website of ASN bank
vacatures friesland	jobs in Friesland (Dutch province)	www.uwv.nl	website of the Employee Insurance Agency
vacatures horeca	vacancies hospitality	zeepbel	bubble
vacatures in de zorg	vacancies in healthcare	zoek werk	search for work

Table 2.B.2: List of Google search terms and their translations/explanations (part 2).

From the set of search terms listed above we discard the Google Trends which have zero values for more than half of the time, before performing the empirical analysis. The final dataset is composed of 182 monthly Google Trends and 173 weekly Google Trends.

2.C Simulation results with non-Gaussian idiosyncratic components

We conduct an additional simulation study in order to assess to what extent the Gaussianity assumptions made on the innovations of the state space model influences the performance of our method. The setting of this additional study is the same as the one discussed in Section 2.4, with the only difference that the nowcast is done in the last period of the sample for 1000 simulation runs. We consider two additional specifications that allow the idiosyncratic components to have distributions that deviate from the Gaussian one, respectively in terms of skeweness and heaviness of the tails:

1. Gaussian-distributed idiosyncratic components:

$$\begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N\left(\mathbf{0}, 0.5\boldsymbol{I}_{n+1}\right).$$

   Notice that this specification is the same as the first one considered in the Section 2.4.

2. Exponentially-distributed idiosyncratic components.

$$\varepsilon_t^{k,y} \sim N\left(0, 0.5\right), \quad \varepsilon_{i,t}^{k,x} \overset{iid}{\sim} \text{Exp}(1), \qquad \text{for } i = 1, \ldots, n.$$

   The exponential distribution is skewed with respect to the Gaussian one.

3. Student's t-distributed idiosyncratic components:

$$\varepsilon_t^{k,y} \sim N\left(0, 0.5\right), \quad \varepsilon_{i,t}^{k,x} \overset{iid}{\sim} t_4, \qquad \text{for } i = 1, \ldots, n.$$

   The t distribution with 4 degrees of freedom has heavier tails with respect to the Gaussian one.

In all specifications $\Lambda \sim U(0,1)$. The generated innovations according to specifications 2 and 3 above, are then standardized to make sure that their distribution is centered around 0 and their variance is equal to 0.5. This ensures that the simulations results are directly comparable with specification 1, and that any deterioration or improvement in the performance of the method can only be attributed to the non-Gaussianity of the innovations.

The additional simulation results are reported in Table 2.C.1 (we report the same measures of nowcast accuracy used in the simulation study of Section 2.4). In terms of MSFE and variance of the nowcasts of the state variables, the distribution does not seem to play a major role. For every specification these two measures improve with a similar magnitude as the correlation parameter increases. The gains are larger for the slope rather than the level of the trend. Their values, relative to the model that does not include any auxiliary series, are broadly lower than one (being around one only when the correlation parameter is small). These results are in line with the ones discussed in Section 2.4. The squared bias, instead, seems to be much more affected by the distribution, as it worsens while deviating from Guassianity and does not improve with a larger correlation parameter. Nonetheless, we notice from Table 2.C.1 that this deterioration of the squared bias has a minor impact on the MSFE since the latter measure is largely composed of its variance component.[12]

We finally look at the consequences of non-Gaussian idiosyncratic components, on the finite-sample distribution of the LR test for the null hypothesis that $\rho = 0$. The formula for computing the LR test is $LR = -2(\mathcal{L}_R - \mathcal{L})$, where $\mathcal{L}_R$ is the value of the log-likelihood under the restriction imposed by the null hypothesis, and $\mathcal{L}$ is the value of the log-likelihood for the unrestricted model, which estimates ρ. We simulate data for each of the three model specifications discussed at the beginning of this section with $\rho = 0$, and we calculate the values of the LR test. We do this for 1000 simulation runs. Notice that under the null hypothesis that $\rho = 0$, and a correct specification of the model, the LR test should be asymptotically χ_1^2-distributed, which (as mentioned in Section 2.5) does not necessarily hold if the model is misspecified, e.g. if the

[12] In particular, we believe that the extremely large values for the relative squared bias are due to the numerical instability caused by the very small absolute values of the squared bias.

true distribution of the error terms is not Gaussian, but a Gaussian distribution is instead used in order to estimate the model. In Figure 2.C.1 we therefore compare the probability densities of the LR tests obtained as described above, to a χ_1^2 distribution. We notice that the density of the LR test is not sensitive to deviations of the idiosyncratic components from Gaussianity. For all three distributions of the error terms considered, i.e., Gaussian, Exponential, and Student's t with 4 degrees of freedom, the density of the LR test is close to a χ_1^2 distribution. These simulation results suggest that our method allows us to conduct inference as usual based on the results of the LR test, even if the distribution of the idiosyncratic components is misspecified.

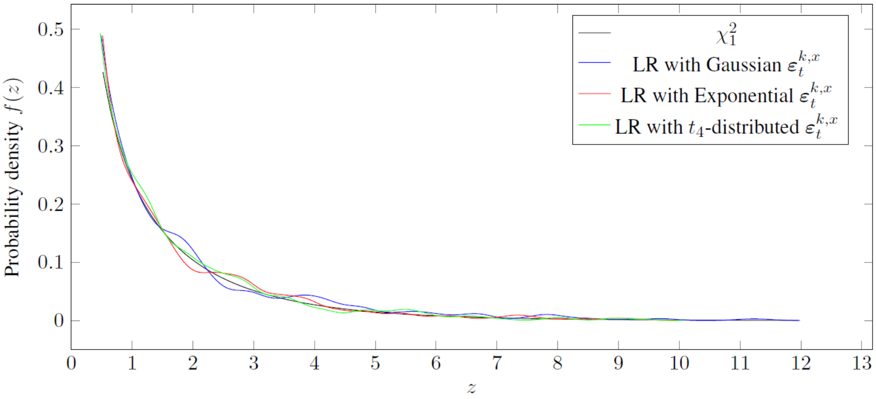


Figure 2.C.1: Probability densities of the LR tests obtained under the null hypothesis that $\rho = 0$, for $n_{\text{sim}} = 1000$ and for the three model specifications discussed at the beginning of Appendix 2.C, respectively, a Gaussian, an Exponential, and a Student's t_4 distribution for the idiosyncratic components, $\varepsilon_t^{k,x}$. We compare these probability densities to a χ_1^2 distribution, which is the asymptotic distribution of the LR test under the null hypothesis, and a correct specification of the model.

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.99$
	Gaussian-distributed idiosyncratic components						
MSFE$(\hat{L}^k_{t\|\Omega_t^-})$	1.016	0.988	0.994	0.941	0.890	0.835	0.773
var$(\hat{L}^k_{t\|\Omega_t^-})$	1.016	0.988	0.994	0.941	0.890	0.835	0.773
bias$^2(\hat{L}^k_{t\|\Omega_t^-})$	0.945	3.805	0.992	0.818	0.608	1.302	1.040
MSFE$(\hat{R}^k_{t\|\Omega_t^-})$	1.048	0.982	0.924	0.754	0.580	0.411	0.253
var$(\hat{R}^k_{t\|\Omega_t^-})$	1.047	0.981	0.924	0.754	0.580	0.411	0.253
bias$^2(\hat{R}^k_{t\|\Omega_t^-})$	1.266	2.152	0.884	2.102	0.795	0.125	0.663
	Exponentially-distributed idiosyncratic components						
MSFE$(\hat{L}^k_{t\|\Omega_t^-})$	1.012	1.004	0.985	0.923	0.880	0.864	0.804
var$(\hat{L}^k_{t\|\Omega_t^-})$	1.012	1.003	0.985	0.923	0.880	0.864	0.804
bias$^2(\hat{L}^k_{t\|\Omega_t^-})$	0.886	1.217	0.124	1.033	551.692	0.267	0.498
MSFE$(\hat{R}^k_{t\|\Omega_t^-})$	1.041	0.998	0.947	0.754	0.547	0.434	0.286
var$(\hat{R}^k_{t\|\Omega_t^-})$	1.042	0.995	0.946	0.749	0.538	0.429	0.276
bias$^2(\hat{R}^k_{t\|\Omega_t^-})$	0.172	1.774	2.579	6.216	390.045	17.565	12.570
	t-distributed idiosyncratic components						
MSFE$(\hat{L}^k_{t\|\Omega_t^-})$	1.011	1.017	0.992	0.970	0.935	0.815	0.788
var$(\hat{L}^k_{t\|\Omega_t^-})$	1.012	1.017	0.992	0.970	0.934	0.815	0.788
bias$^2(\hat{L}^k_{t\|\Omega_t^-})$	0.927	1.233	0.869	13.867	2.116	3.330	7.463
MSFE$(\hat{R}^k_{t\|\Omega_t^-})$	1.039	1.015	0.937	0.782	0.564	0.381	0.231
var$(\hat{R}^k_{t\|\Omega_t^-})$	1.039	1.015	0.937	0.782	0.564	0.380	0.231
bias$^2(\hat{R}^k_{t\|\Omega_t^-})$	0.965	4.126	0.091	0.840	1.119	90.106	53.658

Table 2.C.1: Simulation results from the three settings described in Appendix 2.C. The values are reported relative to the respective measures calculated from the model that does not include the auxiliary series; values < 1 are in favour of our method. $n_{\text{sim}} = 1000$.

2.D Additional empirical results

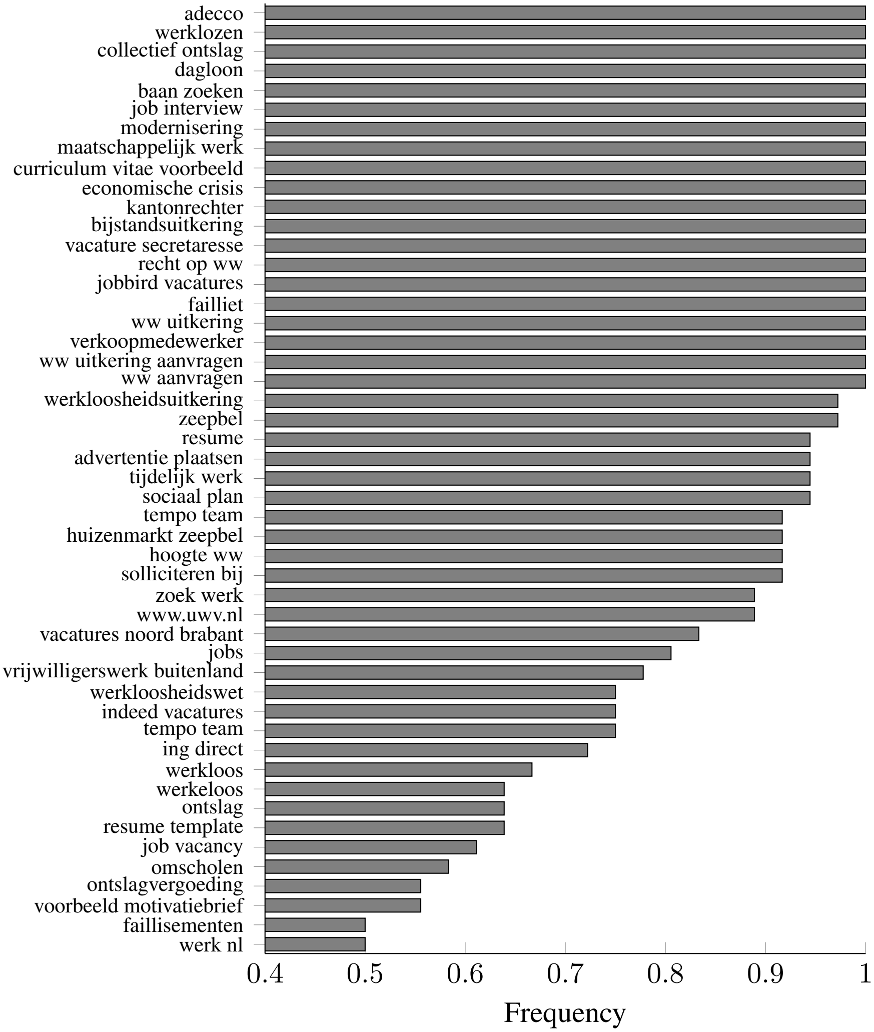


Figure 2.D.1: Frequency of monthly Google search terms selection by the elastic net in the out-of-sample period. A value of 1 means that the variable has been selected in every month of the out-of-sample period. We only report search terms that have been selected at least 50% of the times.

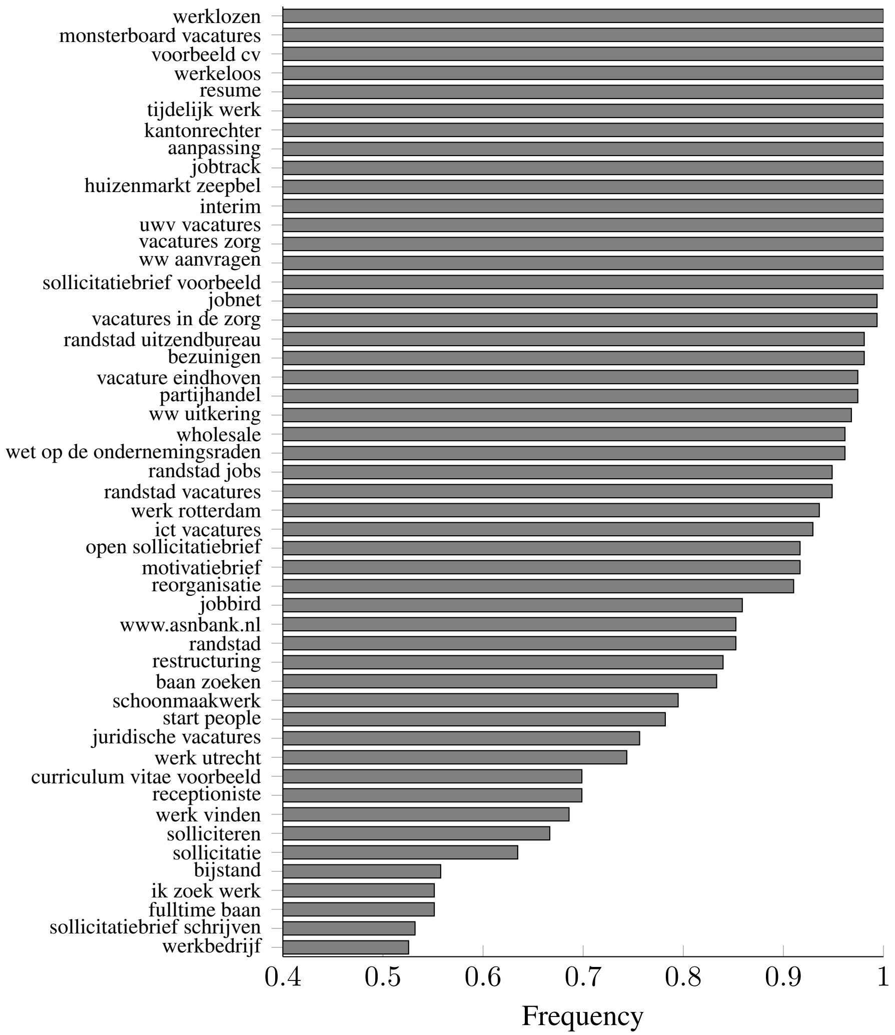


Figure 2.D.2: Frequency of weekly Google search terms (aggregated to the monthly frequency according to equation (2.3.4)) selection by the elastic net in the out-of-sample period. A value of 1 means that the variable has been selected in every week of the out-of-sample period. We only report search terms that have been selected at least 50% of the times.

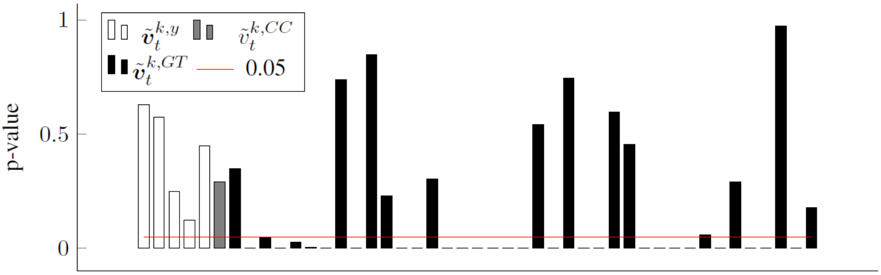


Figure 2.D.3: p-values from the Bowman-Shenton test for individual normality, performed on each of the standardized prediction errors of the labour force, the claimant counts, and the Google Trends series ($\tilde{\boldsymbol{v}}_t^k$). The standardized prediction errors are obtained from the CC & GT model which employs the monthly Google Trends and include two of their factors. The red line represents the confidence level of 0.05.

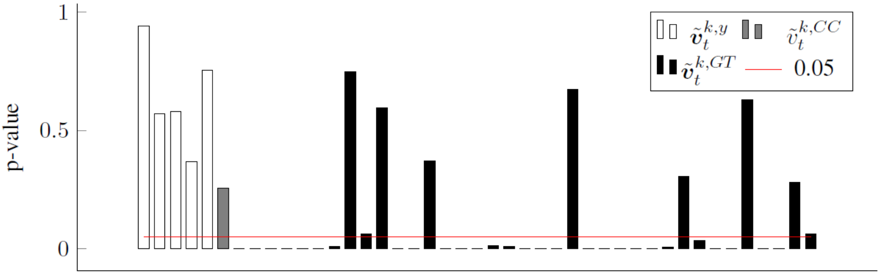


Figure 2.D.4: p-values from the Shapiro-Wilk test for individual normality, performed on each of the standardized prediction errors of the labour force, the claimant counts, and the Google Trends series ($\tilde{\boldsymbol{v}}_t^k$). The standardized prediction errors are obtained from the CC & GT model which employs the weekly Google Trends and include two of their factors, and which iterates the estimation of $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. The red line represents the confidence level of 0.05.

3

Time-varying state correlations in state space models and their estimation via indirect inference

Adapted from: Schiavoni et al. (2021a).

Abstract

Statistics Netherlands uses a state space model to estimate the Dutch unemployment by using monthly series about the labour force surveys (LFS). More accurate estimates of this variable can be obtained by including auxiliary information in the model, such as the univariate administrative series of claimant counts. Legislative changes and economic crises may affect the relation between survey-based and auxiliary series. This time-changing relationship is captured by a time-varying correlation parameter in the covariance matrix of the transition equation's error terms. We treat the latter parameter as a state variable, which makes the state space model become nonlinear and therefore its estimation by Kalman filtering and maximum likelihood infeasible. We therefore propose an indirect inference approach to estimate the static parameters of the model, which employs cubic splines for the auxiliary model, and a bootstrap filter method to estimate the time-varying correlation together with the other state variables of the model. We conduct a Monte Carlo simulation study that shows that our proposed methodology is able to correctly estimate both the time-constant parameters and the state vector of the model. Empirically we find that the financial crisis of 2008 triggered a deeper and more prolonged deviation between the survey-based and the claimant counts series, than a legislative change in 2015. Promptly tackling such changes, which our proposed method does, results in more realistic real-time unemployment estimates.

3.1 Introduction

Official statistics about the labour force, as published by national statistical institutes, are generally based on survey data collected via a rotating panel design in combination with direct estimation procedures, such as the general regression (GREG) estimator (Särndal et al., 1992). Direct or design-based estimators have nice statistical properties under large sample sizes, but the variance of these estimates rapidly becomes unacceptably large in the case of small sample sizes (Rao & Molina, 2015). Small sample sizes typically occur if estimates for short reference periods or estimates at a detailed regional level are required. As a result most national statistical institutes publish quarterly or rolling quarterly figures about the labour force. Since monthly figures are more timely and relevant for policy makers and GREG estimates for monthly labour force figures are not precise enough, Statistic Netherlands implemented a state space model for the production of official monthly labour force figures. This model is used as a form of small area estimation (Rao & Molina, 2015), by using the labour force survey (LFS) data collected over many months to improve the estimates for the current month. The model also accounts for rotation group bias (Bailar, 1975), and serial correlation in the survey errors due to the rotating panel design of the Dutch LFS (Pfeffermann, 1991; van den Brakel & Krieg, 2015).

The aforementioned state space model estimates the Dutch unemployed[1] as an unobserved trend and seasonal component. The state space structure allows us to model not only the unemployment itself, but also its change, its seasonality, the survey errors, and the rotation group bias that affects the observed series from the labour force survey. The estimation accuracy of unemployment can be further improved by augmenting the model with auxiliary series that might be related to it. Harvey and Chung (2000), van den Brakel and Krieg (2016) and Chapter 2 show that the monthly univariate auxiliary series of claimant counts, which is a registry source, can significantly improve the estimation accuracy of the unemployment.

In this chapter we augment the Dutch labour force state space model with the

[1] Throughout the chapter, we use the terms "unemployment", "unemployed" and "unemployed labour force" interchangeably to indicate the total number of unemployed people.

auxiliary series of claimant counts, and we model the relationship between survey-based and auxiliary series as time-varying. This is the main novelty of the chapter. Modeling relationships between variables as time-varying allows us to tackle changes that are triggered, for instance, by economic crises. Time-changing relations between different sources of data that aim at measuring the same variables, can also be due to legislative changes. Of course these events are of different nature and can result in different evolutions over time for the relationship of interest. This one can, for instance, show an abrupt change and then be followed by a gradual recovery, or be just characterised by a smooth adjustment to a new level. In a Monte Carlo simulation study we will therefore look at different possible patterns of the relationship between two time series.

We let the time-varying relation be captured by a time-varying correlation in the covariance matrix of the disturbances of the state space model's transition equation. Solving our problem, i.e., extracting this time-varying state correlation, is therefore already challenging by the fact that this parameter relates innovations of components that are unobserved. Additionally, data and their respective log-likelihood functions, are much less informative about correlations than other parameters, such as means or variances.

We treat the time-varying correlation that enters this covariance matrix as an additional state variable with its own dynamic equation, which makes this parameter random, i.e. subject to its own source of error. The state space model therefore becomes nonlinear. The nonlinearity of the model makes it infeasible the estimation of the state variables by standard Kalman filtering, which in turns makes the likelihood function (needed to estimate the static parameters of the model) intractable. There are several frequentist methods that can be employed in order to estimate nonlinear state space models. Importance sampling (see Jungbacker and Koopman (2007), Koopman et al. (2015), and Koopman et al. (2018)) and the Extended Kalman filter (explained in Durbin and Koopman (2012, Chapter 10)) are some examples, and they allow us to estimate both the state variables and the time-constant parameters of the nonlinear model. The reprojection method of Gallant and Tauchen (1998) is another technique that can be employed in order to estimate the state vector but not the static parameters of nonlinear state space models. Nonetheless, the implementation of these methods in our setting is hampered by the nonlinearity being in the transition equation and involving not only state variables but also disturbance terms (as will become clear at the beginning of Section 3.3). The above-mentioned methods are more suited to solve nonlinearities appearing in the observation equation or involving only state variables. Shifting our nonlinear problem from the transition to the observation equation is non-trivial. Alternatively, sequential Monte Carlo methods, such as particle filtering, can be employed (and we do) for the estimation of the state vector, but the resulting log-likelihood function is not continuous with respect to the static parameters of the nonlinear model, which hinders their estimation via this approach (Creal, 2012). We circumvent this problem by estimating the time-constant parameters by indirect inference, before applying particle filtering.

The indirect inference method, originally proposed by Gourieroux et al. (1993), requires the use of an auxiliary model which approximates the true one, but which can also be easily estimated. In our case the auxiliary model employs a deterministic specification for the time-varying correlation, which preserves the linearity of the state space model. This deterministic specification assumes a smooth change over time for the correlation, and is based on the cubic splines estimation method. The latter approach requires the change points in time to be chosen a priori, and can already be employed, by itself, in estimating the time-varying state correlation. Koopman et al. (2006) and Proietti and Hillebrand (2017) already utilised this method for modeling time-varying parameters in state space models. However, the use of cubic splines for the auxiliary model in the indirect inference estimation of static parameters, not only in state space models, has not been explored before. Once the time-constant parameters have been estimated by indirect inference, we employ a Rao-Blackwellised (Chen & Liu, 2000) bootstrap filter (Gordon et al., 1993), which is a type of sequential Monte Carlo algorithm, and more specifically a type of particle filter, in order to estimate the state variables of the model. The Rao-Blackwellisation of the bootstrap filter is needed in order to simplify the estimation of the state vector when this is large, which is the case in the Dutch labour force model extended with the claimant counts series.

Monfardini (1998) and Gagliardini et al. (2017) are works very close to this chapter as they both employ indirect inference in order to estimate the static parameters of nonlinear state space models, where the nonlinearity arises from stochastic variances in the innovations of the observation equation. In the former paper Autoregressive and Moving Average (ARMA) are employed as auxiliary models, whereas the latter work makes use of Mixed Data Sampling (MIDAS) regressions and Autoregressive Conditional Heteroskedasticity (ARCH) models, and it also provides a filtering step for the estimation of state variables based on the reprojection method of Gallant and Tauchen (1998). Stochastic variances are a rather common source of nonlinearity in state space models and there are many, also applied, papers that deal with them. Stock and Watson (2007) and Antolin-Diaz et al. (2017) are two examples, and they consider nonlinear state space models, respectively for US inflation and a set of US macroeconomic variables, where both the innovations in the measurement and transition equations have stochastic variances. The models are estimated via the Bayesian approach of Markov Chain Monte Carlo (MCMC), which could probably also be employed to estimate our nonlinear model, but we do not venture into Bayesian techniques. Both papers show how recession periods can play a big role in triggering changes in parameters, in their cases by boosting volatilities. To the best of our knowledge, this chapter is the first one to deal with the estimation of a stochastic state correlation, and that therefore employs particle filtering for this purpose.

We conduct a Monte Carlo simulation study in order to assess the performance of our proposed estimation methods: the one purely based on cubic splines, and the one which combines indirect inference and Rao-Blackwellised bootstrap filtering. We investigate not only how they are able to estimate the time-constant parameters of the model as well as the true time-varying relation, but also to what extent estimating this relation as time-varying, instead of time-constant, yields gains in the estimation accuracy of unobserved components of interest. We then conduct an empirical study to estimate the Dutch unemployment by means of the labour force state space model augmented with the auxiliary series of claimant counts, while modeling the relationship between survey-based and auxiliary series as time-varying, with our proposed methods.

The chapter is structured as follows. Section 3.2 presents a description of the state space model used by Statistics Netherlands for estimating monthly unemployment figures, and its extension with the univariate auxiliary series of claimant counts. Section 3.3 describes our proposed methods for the estimation of the time-varying correlation, together with the remaining state variables and the static parameters of the nonlinear state space model. Sections 3.4 and 3.5 report the results of, respectively, the Monte Carlo simulation study and the empirical application. Finally, Section 3.6 concludes the chapter.

3.2 The Dutch labour force model and its extension

The Dutch LFS is conducted according to a rotating panel design. Each month a new sample, drawn according to the stratified two-stage cluster design described in van den Brakel and Krieg (2015) and Chapter 2, enters the panel and is interviewed five times at quarterly intervals. After the fifth interview, the sample leaves the panel. The sample that is interviewed for the j^{th} time is called the j^{th} wave of the panel, for $j = 1, ..., 5$. This rotation design implies that in each month five samples are observed, which over time generate a five-dimensional time series of the survey-based unemployed labour force, defined as population total (see Table 2.1 for a visualisation of the rotation panel design of the Dutch LFS).

Let $y_{j,t}$ denote the general regression (GREG) estimate (Särndal et al., 1992) for the Dutch unemployment in month t based on the sample observed in wave j. Now $\boldsymbol{y}_t = (y_{1,t}, \dots, y_{5,t})'$ denotes the vector with the five GREG estimates for the Dutch unemployment in month t. This five-dimensional vector of GREG estimates is cast in a state space model whose measurement equation takes the expression:

$$\boldsymbol{y}_t = \boldsymbol{\imath}_5 \theta_{y,t} + \boldsymbol{\lambda}_t + \boldsymbol{e}_t, \qquad (3.2.1)$$

where $\boldsymbol{\imath}_5$ is a five-dimensional column vector of ones, and $\theta_{y,t}$ is a common unobserved state variable among the five-dimensional waves of the survey-based unemployed labour force, and it represents the Dutch unemployment itself. As such, it is our variable of interest, and as an unobserved component, it is assumed to be unknown and estimable. The reason why the unemployment is re-estimated by means of a state space model, using the GREG estimates as observed series, is because the latter are considered too volatile to produce sufficiently reliable monthly estimates for the unemployed labour force at monthly frequency. The additional estimation method via state space models, which was originally proposed by Pfeffermann (1991), improves the precision of the monthly estimates for the unemployment with sample information from previous periods, and can therefore be seen as a form of small area estimation (Rao & Molina, 2015).

The state variable of interest, $\theta_{y,t}$, is assumed to be composed of a trend and a seasonal component (which means that strictly speaking $\theta_{y,t}$ is the sum of state variables, but for simplicity we refer to it as state variable throughout the chapter):

$$\theta_{y,t} = L_{y,t} + S_{y,t}.$$

The exclusion of an innovation term in the formula above is motivated by Bollineni-Balabay et al. (2017). The transition equations for the level $(L_{y,t})$ and the slope $(R_{y,t})$ of the trend are, respectively:

$$\begin{aligned} L_{y,t+1} &= L_{y,t} + R_{y,t}, \\ R_{y,t+1} &= R_{y,t} + \eta_{R,y,t}, \quad \eta_{R,y,t} \sim N\left(0, \sigma^2_{R,y}\right). \end{aligned}$$

The random walk specification for the slope means that the latter is assumed to be integrated of order 1, I(1), which implies that its first differences are assumed to be stationary (i.e., mean-reverting). Consequently, the level of the trend, which is expressed as the cumulative sum of the slope, is integrated of order 2, I(2), implying that its second differences are stationary. The absence of an innovation term for the trend's level is motivated in van den Brakel and Krieg (2016) as being the result of Likelihood Ratio testing, and implies a smoothness assumption on the level of the trend in unemployment.

The trigonometric stochastic seasonal component allows for the seasonality to vary over time, and it is modeled as in Durbin and Koopman (2012, Chapter 3):

$$
\begin{aligned}
&S_{y,t} = \sum_{l=1}^{6} S_{1,y,l,t}, \\
&\begin{pmatrix} S_{1,y,l,t+1} \\ S_{2,y,l,t+1} \end{pmatrix} = \begin{bmatrix} \cos(h_l) & \sin(h_l) \\ -\sin(h_l) & \cos(h_l) \end{bmatrix} \begin{pmatrix} S_{1,y,l,t} \\ S_{2,y,l,t} \end{pmatrix} + \begin{pmatrix} \eta_{1,\omega,y,l,t} \\ \eta_{2,\omega,y,l,t} \end{pmatrix}, \\
&\begin{pmatrix} \eta_{1,\omega,y,l,t} \\ \eta_{2,\omega,y,l,t} \end{pmatrix} \sim N\left(\mathbf{0}, \sigma^2_{\omega,y} \boldsymbol{I}_2\right),
\end{aligned}
$$

where $h_l = \frac{\pi l}{6}$, for $l = 1, \ldots, 6$, and $\boldsymbol{I}_2$ is a 2×2 identity matrix.

Rotating panel designs can induce Rotation Group Bias (RGB), i.e., systematic differences among the observations in the subsequent waves (Bailar, 1975). van den Brakel and Krieg (2015) argue that, for the Dutch LFS, the estimates for unemployment based on the first wave are systematically larger than the estimates based on the follow-up waves. Some of the reasons that trigger this phenomenon are discussed in Chapter 2, and they suggest that the answers from the first wave of interviews have to be considered as being the most reliable ones and not to be affected by the RGB. The five-dimensional state vector $\boldsymbol{\lambda}_t$ in equation (3.2.1) accounts for the RGB in the second to fifth wave, as proposed in Pfeffermann (1991), and its last four elements are modeled as a random walk because they are supposed to capture the time-dependent differences with respect to the first wave:

$$
\begin{aligned}
&\lambda_{1,t+1} = 0, \\
&\lambda_{j,t+1} = \lambda_{j,t} + \eta_{\lambda,j,t}, \quad \eta_{\lambda,j,t} \sim N\left(0, \sigma^2_{\lambda}\right), \quad j = 2, \ldots, 5.
\end{aligned}
$$

Notice that $\lambda_{1,t+1} = 0$ because it is assumed that the first wave is not affected by the RGB.

The rotating panel design also induces autocorrelation among the survey errors in the follow-up waves. In order to account for this autocorrelation, the survey errors, which are represented by the five-dimensional vector $\boldsymbol{e}_t$ in equation (3.2.1), are treated as state variables. The transition equation for the survey errors takes the following form:

$$
\begin{aligned}
& e_{j,t} = c_{j,t}\xi_{j,t}, \quad c_{j,t} = \sqrt{\text{var}\left(y_{j,t}\right)}, \quad j = 1, \ldots, 5, \\
& \xi_{1,t+1} \sim N\left(0, \sigma^2_{\nu_1}\right), \\
& \xi_{j,t+1} = \delta\xi_{j-1,t-2} + \nu_{j,t}, \quad \nu_{j,t} \sim N\left(0, \sigma^2_{\nu_j}\right), \quad j = 2, \ldots, 5, \quad |\delta| < 1, \\
& \ \text{var}\left(\xi_{j,t}\right) = \sigma^2_{\nu_j} / \left(1 - \delta^2\right), \quad j = 2, \ldots, 5.
\end{aligned}
$$

The survey errors of all waves, $e_{j,t}$, are assumed to be proportional to the standard errors of the GREG estimates, $\sqrt{\text{var}\left(y_{j,t}\right)}$, for $j = 1, \ldots, 5$, in order to account for heterogeneity in their variances, which are caused by, for instance, changing sample sizes over time. The scaled sampling errors $\xi_{j,t}$, $j = 1, \ldots, 5$, capture the serial autocorrelation induced by the sampling overlap of the rotating panel. Since in the first wave of interview samples are observed for the first time, the survey errors of the first wave are not autocorrelated with survey errors of previous periods. The survey errors of the second to fifth wave are, instead, correlated with the survey errors of the previous wave three months before. For this reason van den Brakel and Krieg (2009), following an approach proposed by Pfeffermann et al. (1998), suggest to model the survey errors with an auto-regressive process of order 3, AR(3), without including the first and second lag.

In order to achieve more accurate estimates, or forecasts, of the unemployment, it is possible to augment the model with auxiliary series that are related to this variable. Harvey and Chung (2000), van den Brakel and Krieg (2016) and Chapter 2 show that including in the model the univariate auxiliary series of monthly claimant counts, which represents the number of people claiming unemployment benefits and which is a registry source, can significantly improve the accuracy of estimation (and nowcasting) of unemployment. Only unemployed people who have worked enough time, and therefore paid enough taxes, can receive unemployment benefits in the Netherlands, for a maximum of three years and two months, whether they are employed or not at the end of this period. The claimant counts therefore tend to underestimate the Dutch long-term unemployment. Figure 3.2.1 displays the monthly time series of the GREG estimates and the claimant counts, from January 2004 until March 2020. It is possible to see how the two series overall tend to follow the same trend over time, but deviate from each other between 2010 and 2016.

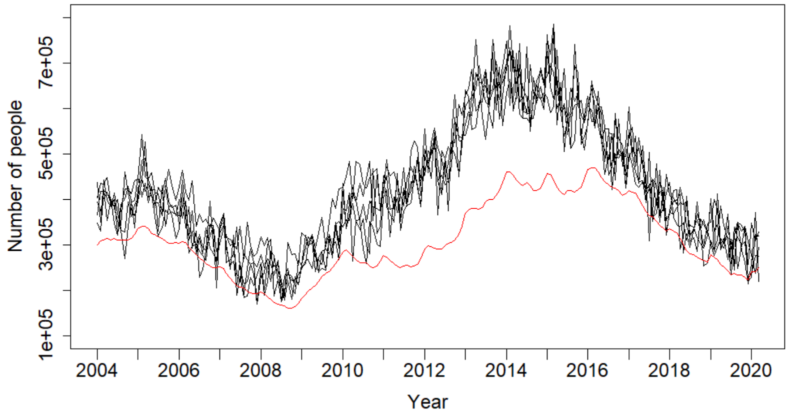


Figure 3.2.1: Monthly GREG estimates for Dutch unemployment based on the labour force survey (black), $\boldsymbol{y}_t$, and the univariate time series of claimant counts in the Netherlands (red).

If we let $x_{CC,t}$ be the univariate series of monthly claimant counts, the labour force model augmented with this auxiliary series, looks as follows:

$$\begin{pmatrix} \boldsymbol{y}_t \\ x_{CC,t} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\imath}_5\theta_{y,t} \\ \theta_{CC,t} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t \\ 0 \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_t \\ \varepsilon_{CC,t} \end{pmatrix}, \quad \varepsilon_{CC,t} \sim N\left(0, \sigma^2_{\varepsilon,CC}\right), \tag{3.2.2}$$

$$\begin{pmatrix} \theta_{y,t} \\ \theta_{CC,t} \end{pmatrix} = \begin{pmatrix} L_{y,t} \\ L_{CC,t} \end{pmatrix} + \begin{pmatrix} S_{y,t} \\ S_{CC,t} \end{pmatrix}, \tag{3.2.3}$$

$$\begin{pmatrix} L_{y,t+1} \\ L_{CC,t+1} \end{pmatrix} = \begin{pmatrix} L_{y,t} \\ L_{CC,t} \end{pmatrix} + \begin{pmatrix} R_{y,t} \\ R_{CC,t} \end{pmatrix}, \tag{3.2.4}$$

$$\begin{pmatrix} R_{y,t+1} \\ R_{CC,t+1} \end{pmatrix} = \begin{pmatrix} R_{y,t} \\ R_{CC,t} \end{pmatrix} + \begin{pmatrix} \eta_{R,y,t} \\ \eta_{R,CC,t} \end{pmatrix}, \tag{3.2.5}$$

with

$$\operatorname{cov}\begin{pmatrix} \eta_{y,R,t} \\ \eta_{CC,R,t} \end{pmatrix} = \begin{bmatrix} \sigma^2_{R,y} & \rho\sigma_{R,y}\sigma_{R,CC} \\ \rho\sigma_{R,y}\sigma_{R,CC} & \sigma^2_{R,CC} \end{bmatrix}. \tag{3.2.6}$$

The augmented state space model above shows that the series of claimant counts is also supposed to be composed of a level and a seasonal component, which are assumed to have the same transition equations of the level and the seasonal components of $\theta_{y,t}$.

The static parameters (which are also referred to as “hyperparameters” in the state space literature) and the state variables of the model defined by equations (3.2.2)-(3.2.6) are estimated, respectively, by maximum likelihood using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimisation algorithm, and by the Kalman filter (details about these estimation methods are provided in Section 3.3). A diffuse initialisation of the Kalman filter is used for all state variables of the model, except for the 13 state variables that define the autocorrelation structure of the survey errors, for which we use the exact initialisation of Bollineni-Balabay et al. (2017).

The transition equation (3.2.5) implies that the trend’s slopes of the survey-based and the claimant counts series have the same order of integration: they are both I(1). Their innovations are allowed to be correlated, as their covariance matrix in equation (3.2.6) shows. Harvey and Chung (2000) show, via simulations, that if the magnitude of this correlation parameter is large, there are gains in the accuracy of estimation and nowcast, respectively in terms of MSE and MSFE, of the Kalman filter estimators of $\theta_{y,t}$, $L_{y,t}$ and $R_{y,t}$. In the special case where the absolute value of the correlation parameter is equal to 1, the covariance matrix of equation (3.2.6) is not full rank any more, which means that the corresponding state variables have the same source of error, and are said to be cointegrated. This correlation parameter is of key importance in our study, because it represents the means for the Dutch labour force model to exploit auxiliary information.

Notice that the smooth trend model specification for the claimant counts’ trend implies that the claimant counts series is I(2).

We employ the method proposed by Harvey and Chung (2000) in order to incorporate auxiliary information in the model, by augmenting the observed series with the auxiliary one. Alternatively, the claimant counts series could be added as a regressor in the measurement equation. However, with this latter option the main part of the trend in unemployment would be explained by the auxiliary series, and its filtered estimates would contain a residual trend instead of the unemployment's trend. Since trend estimates are published as part of the monthly official labour force figures, this approach is not an option for Statistics Netherlands, and we do not consider it in this chapter.

Model (3.2.1) has been used by Statistics Netherlands since 2010 for the production of official labour force figures. To further improve the precision of the time series model estimates, Statistics Netherlands decided in 2014 to augment this model with the series of Dutch claimant counts. Until the end of 2014, survey-based and auxiliary series were actually cointegrated. In February 2015, some changes were implemented in the registration of claimant counts. Namely, since then people that find a job can receive unemployment benefits up to two months after having found the job. These people are therefore still part of the claimant counts for those two additional months despite being actually employed. This resulted in systematic higher time series estimates for the unemployed labour force in the second quarter of 2015, compared to the model without claimant counts. It was anticipated that this legislative change would disturb the relationship between the survey-based and the claimant counts series, and that a model assuming a time constant correlation incorrectly would not not observe the drop in this parameter during the first months after the legislative change. As a result, Statistics Netherlands went back to the model without claimant counts in June 2015, and revised the monthly figures that were published from March until May 2015. This model has been used for publishing official monthly figures about the Dutch labour force since then. Hence the main purpose of this chapter to model the relationship between survey-based and auxiliary series as time-varying, by proposing a method for the estimation of time-varying state correlations in state space models. The correlation parameter in the covariance matrix of equation (3.2.6) is therefore, hereinafter, assumed to be time-varying.

3.3 Method for the estimation of a time-varying state correlation

This section describes our proposed approach to model a time-varying state correlation, when this is assumed to be random. For the sake of generality, the auxiliary series included in the model will be indicated as $\boldsymbol{x}_t$ (which in the case of the claimant counts is just a scalar). Additionally, we need to impose a restriction on the space of ρ_t, namely of being between -1 and 1. We therefore re-parametrise the correlation parameter as

$$\rho_t = \tanh\left(\gamma_t\right),$$

for $t = 1, \ldots, T$, and where $\tanh$ is the hyperbolic tangent, which is a time-invariant, continuous, invertible, and twice differentiable function.

We let $\boldsymbol{z}_t = (\boldsymbol{y}_t', \boldsymbol{x}_t')'$ be the $n \times 1$ vector that collects all the observed series, and $Z_{t-1} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{t-1}\}$ be the available information set at time t. The Dutch labour force model (3.2.2)-(3.2.6) can be compactly written as

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{Z}\boldsymbol{\alpha}_t + \boldsymbol{M}\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N\left(\boldsymbol{0}, \boldsymbol{H}\right) \\ \boldsymbol{\alpha}_{t+1} &= \boldsymbol{T}\boldsymbol{\alpha}_t + \boldsymbol{R}\boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N\left(\boldsymbol{0}, \boldsymbol{Q}_t\right), \end{aligned} \tag{3.3.1}$$

for $t = 1, \ldots, T$, where T is the sample size, $\boldsymbol{\alpha}_t$ is the $p \times 1$ vector of state variables and $\boldsymbol{\eta}_t$ is the corresponding $r \times 1$ vector of disturbances. The respective covariance matrix, $\boldsymbol{Q}_t$, is varying over time because it contains the time-varying correlation parameter, ρ_t. The $p \times p$ matrix $\boldsymbol{T}$ defines the dynamic structure of the state variables, and $\boldsymbol{R}$ is a $p \times r$ selection matrix. The $n \times p$ matrix $\boldsymbol{Z}$ links the observed series to the state vector, $\boldsymbol{\varepsilon}_t$ is the $q \times 1$ disturbance vector of the observation equation, with covariance matrix $\boldsymbol{H}$, and $\boldsymbol{M}$ is a $n \times q$ selection matrix. We assume the matrices $\boldsymbol{Z}$, $\boldsymbol{M}$, $\boldsymbol{H}$, $\boldsymbol{T}$ and $\boldsymbol{R}$ to be non-stochastic[2]. If we also assume $\boldsymbol{Q}_t$ to be non-stochastic, which happens if ρ_t is actually constant, or if we assume a deterministic specification

[2] The exact expressions for all these matrices in the extended Dutch labour force model, can be found in Chapter 2.

that captures the time-variation of ρ_t, then the state space model (3.3.1) is linear. We indicate with $\boldsymbol{\beta}$ the vector of unknown time-constant parameters of the linear model (which are contained in the above-listed matrices). Then, conditionally on the information set and $\boldsymbol{\beta}$, the observations and the state vector are Gaussian: $(\boldsymbol{z}_t|Z_{t-1};\boldsymbol{\beta}) \sim N(\boldsymbol{Z}\boldsymbol{a}_t, \boldsymbol{F}_t)$ and $(\boldsymbol{\alpha}_t|Z_{t-1};\boldsymbol{\beta}) \sim N(\boldsymbol{a}_t, \boldsymbol{P}_t)$. Therefore, the log-likelihood function for $\boldsymbol{z}_t$ takes the form[3]

$$\ell = \sum_{t=1}^{T} \ell_t, \tag{3.3.2}$$

with

$$\ell_t = \log p(\boldsymbol{z}_t|Z_{t-1};\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det \boldsymbol{F}_t) - \frac{1}{2}\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t, \tag{3.3.3}$$

for $t = 1, \dots, T$, where the prediction error $\boldsymbol{v}_t$ and its covariance matrix $\boldsymbol{F}_t$ are evaluated by the following Kalman Filter recursions:

$$\begin{aligned} \boldsymbol{v}_t &= \boldsymbol{z}_t - \boldsymbol{Z}\boldsymbol{a}_t & \boldsymbol{P}_{t|t} &= \boldsymbol{P}_t - \boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\boldsymbol{Z}\boldsymbol{P}_t \\ \boldsymbol{F}_t &= \boldsymbol{Z}\boldsymbol{P}_t\boldsymbol{Z}' + \boldsymbol{M}\boldsymbol{H}\boldsymbol{M}' & \boldsymbol{a}_{t+1} &= \boldsymbol{T}\boldsymbol{a}_t + \boldsymbol{K}_t\boldsymbol{v}_t \\ \boldsymbol{K}_t &= \boldsymbol{T}\boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1} & \boldsymbol{P}_{t+1} &= \boldsymbol{T}\boldsymbol{P}_t(\boldsymbol{T} - \boldsymbol{K}_t\boldsymbol{Z})' + \boldsymbol{R}\boldsymbol{Q}_t\boldsymbol{R}', \\ \boldsymbol{a}_{t|t} &= \boldsymbol{a}_t + \boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t & & \end{aligned} \tag{3.3.4}$$

[3] In case of a diffuse initialisation of the Kalman filter, we employ the following diffuse log-likelihood (Harvey, 1989), instead of equations (3.3.2)-(3.3.3):

$$\ell_d = -\frac{Tn}{2}\log(2\pi) - \frac{1}{2}\sum_{t=d+1}^{T}\left[\log(\det \boldsymbol{F}_t) - \frac{1}{2}\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right],$$

with d being the number of nonstationary state variables of the target observed series (in the empirical application of Section 3.5, d is the number of unobserved components of the GREG estimates for which a diffuse initialisation is used).

for $t = 1, \dots, T$. The vector $\boldsymbol{a}_{t|t}$ represents the filtered estimate of $\boldsymbol{\alpha}_t$, and $\boldsymbol{P}_{t|t}$ is its estimated covariance matrix. The one-step-ahead prediction of the state vector, together with its predicted covariance matrix, are represented, respectively, by $\boldsymbol{a}_{t+1}$ and $\boldsymbol{P}_{t+1}$. Assuming normality of the innovations is standard in state space models because it allows us to estimate $\boldsymbol{\beta}$ by maximising the log-likelihood (3.3.2). Under the normality assumption, the Kalman filter yields the minimum variance unbiased estimator of the state vector. However, if the Gaussianity assumption of the innovations is not met, then the Kalman filter still provides the minimum variance *linear* unbiased estimator of the state vector, as long as the model is linear (Durbin & Koopman, 2012, Chapter 4).

The correlation parameter, ρ_t, can be assumed to be stochastic, instead of deterministic, by treating γ_t as an additional state variable with its own transition equation:

$$\gamma_{t+1} = \gamma_t + \eta_{\gamma,t}, \quad \eta_{\gamma,t} \sim N\left(0, \sigma_\gamma^2\right) \tag{3.3.5}$$

for $t = 1, \dots, T$. The state vector hence becomes $\boldsymbol{\alpha}_t^* = (\boldsymbol{\alpha}_t', \gamma_t)'$, yielding what we refer to as the “nonlinear state space model”:

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{Z}^* \boldsymbol{\alpha}_t^* + \boldsymbol{M}\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N\left(\boldsymbol{0}, \boldsymbol{H}\right) \\ \boldsymbol{\alpha}_{t+1}^* &= \boldsymbol{T}^* \boldsymbol{\alpha}_t^* + \boldsymbol{R}^* \boldsymbol{\eta}_t^*, \quad \boldsymbol{\eta}_t^* \sim N\left(\boldsymbol{0}, \boldsymbol{Q}_t^*\right), \end{aligned} \tag{3.3.6}$$

for $t = 1, \dots, T$, where $\boldsymbol{\eta}_t^* = (\boldsymbol{\eta}_t', \eta_{\gamma,t})'$, and $\boldsymbol{Z}^*$, $\boldsymbol{T}^*$, $\boldsymbol{R}^*$ and $\boldsymbol{Q}_t^*$ are straightforwardly obtained by adding γ_t as an additional state variable and $\eta_{\gamma,t}$ as an additional state innovation. We let $\boldsymbol{\tau}$ be the parameter vector that collects all the static parameters of the nonlinear model, i.e., the time-constant parameters appearing in $\boldsymbol{H}$, $\boldsymbol{T}^*$, and $\boldsymbol{Q}_t^*$ (including σ_γ) of model (3.3.6). Notice that the matrices $\boldsymbol{M}$ and $\boldsymbol{H}$ are the same as the ones specified for the linear state space model (3.3.1). Model (3.3.6) is nonlinear in its transition equation. To see this, let us define with $\boldsymbol{C}_{Q,t}$ the Cholesky decomposition of $\boldsymbol{Q}_t^*$ (i.e., such that $\boldsymbol{C}_{Q,t}\boldsymbol{C}_{Q,t}' = \boldsymbol{Q}_t^*$). Then $\boldsymbol{\eta}_t^* \sim N\left(\boldsymbol{0}, \boldsymbol{Q}_t^*\right)$ can be re-written as $\boldsymbol{C}_{Q,t}\boldsymbol{\eta}_t^* \sim N\left(\boldsymbol{0}, \boldsymbol{I}_{r+1}\right)$, where $\boldsymbol{C}_{Q,t}\boldsymbol{\eta}_t^*$ involves multiplications (i.e., nonlinear structures) of random components (ρ_t and $\boldsymbol{\eta}_t$). This implies that the state vector of the nonlinear model, $\boldsymbol{\alpha}_t^*$, cannot be estimated by the standard

Kalman filter recursions (3.3.4), and the consequent evaluation of the log-likelihood (3.3.2) is therefore precluded. As already mentioned in the Introduction, we solve this problem by proposing an indirect inference approach, which makes use of cubic splines for the auxiliary model, in order to estimate $\boldsymbol{\tau}$, and then estimating the state vector, $\boldsymbol{\alpha}_t^*$, with the Rao-Blackwellised Bootstrap filter (RBBF). The next two sub-sections explain our proposed methodology in detail.

3.3.1 Indirect inference with a cubic splines approximate model

When the likelihood function needed to estimate the static parameters of a model is intractable, which, as explained above, is the case for a nonlinear state space model, Gourieroux et al. (1993) propose to indirectly estimate these parameters via the optimisation of an "incorrect criterion". This can be, for instance, the exact log-likelihood of an approximate model, which can therefore be easily estimated. In our case a natural approximate model would be a linear Gaussian state space model, which assumes a deterministic specification for the time-varying correlation. We do so by means of cubic splines.

Cubic splines are continuous piecewise polynomial functions of cubic order whose function values and first and second derivatives agree at the points where they join (Smith, 2008). The abscissas of these joint points are called knots and are chosen to determine the complexity of the approximation. A common choice is to set the knots to evenly partition the support $1, \ldots, T$ of t. Specifically, in the Monte Carlo simulation and empirical studies, we specify two choices for the location of the knots: at the quartiles and septiles of the sample size. We then let an information criterion decide on the best choice of the knots' locations[4]. If the date of a structural change in the time-varying

[4] The information criteria that we employ are the Akaike and the Bayesian, respectively given by the following expressions (Durbin & Koopman, 2012, Chapter 7):

$$\begin{aligned}\text{AIC} &= \frac{1}{T}\left[-2\ell_d + 2\left(d + \dim(\boldsymbol{\beta})\right)\right]\\ \text{BIC} &= \frac{1}{T}\left[-2\ell_d + \log\left(T\right)\left(d + \dim(\boldsymbol{\beta})\right)\right],\end{aligned}$$

parameter is known, it can be used as (additional) knot. The polynomial order controls for the smoothness of the splines and is treated as fixed (Hansen, 2019).

The deterministic time-varying specification for γ_t, based on cubic splines, is

$$\gamma_t = \boldsymbol{w}_t' \boldsymbol{\phi}, \tag{3.3.7}$$

where $\boldsymbol{w}_t$ is a $k \times 1$ vector of known weights, which depends on t, on the position of the knots and the distance between them; it corresponds to the t^{th} row of a $T \times k$ matrix $\boldsymbol{W}$ of splines weights constructed as in Poirier (1973), where k is equal to the number of knots. Moreover, $\boldsymbol{\phi}$ is a $k \times 1$ vector of coefficients. Each element ϕ_j of $\boldsymbol{\phi}$, represents the value of γ_t at the j^{th} knot, for $j = 1, \ldots, k$. We use a natural cubic spline with truncated power polynomial basis. The natural feature restricts the spline functions to be linear beyond the boundary knots, in order to decrease the large variance that affects cubic splines at the boundaries, while paying the price for a larger bias (Hastie et al., 2009, Chapter 5). Notice that the weights $\boldsymbol{w}_t$ make sure that γ_t varies over time, but they are all pre-specified, which means that there is no stochastic component in specification (3.3.7). This makes sure that the state space model remains linear. Equations (3.3.1) and (3.3.7) define what we refer to as the "approximate/auxiliary model" or the "cubic splines model". The parameter vector $\boldsymbol{\beta}$ of this approximate linear model therefore contains the time-constant parameters appearing in $\boldsymbol{H}$, $\boldsymbol{T}$, and $\boldsymbol{Q}_t$ of model (3.3.1), together with $\boldsymbol{\phi}$. Vectors $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ only differ because the latter contains σ_γ instead of $\boldsymbol{\phi}$. Very importantly, $\boldsymbol{\beta}$ is easily estimated by maximising the log-likelihood (3.3.2). We indicate with $\hat{\boldsymbol{\beta}}$ the maximum likelihood estimate of $\boldsymbol{\beta}$. The cubic splines estimate for ρ_t is therefore equal to $\hat{\rho}_{\text{CS},t} = \tanh(\boldsymbol{w}_t' \hat{\boldsymbol{\phi}})$.

In the context of state space models, a cubic splines approach to estimate time-varying parameters has already been employed by Koopman et al. (2006) and Proietti and Hillebrand (2017). They use this method to model, respectively,

where $\dim(\boldsymbol{\beta})$ is the dimension of $\boldsymbol{\beta}$. We use the former criterion in the Monte Carlo simulation study of Section 3.4, and the latter in the empirical application of Section 3.5. This is because BIC is notoriously more parsimonious than AIC, an it therefore prevents an unnecessary increase in the (already high) complexity of the Dutch labour force model.

time-changing volatilities and smoothly changing parameters over the seasons of the year.

After $\boldsymbol{\beta}$ has been estimated by maximum likelihood on the observed series, Z_T, the indirect inference approach proceeds by simulating, for a given value of $\boldsymbol{\tau}$, S paths of Z_T according to the nonlinear model. The nonlinear model is therefore required to be simulable (which is the case for model (3.3.6)). The cubic splines model is re-estimated on every simulated path $Z_T^{(s)}(\boldsymbol{\tau})$, for $s = 1, \ldots, S$, yielding S different maximum likelihood estimates, $\hat{\boldsymbol{\beta}}^{(s)}(\boldsymbol{\tau})$. The parameter vector of the nonlinear model, $\boldsymbol{\tau}$, is then estimated as

$$\hat{\boldsymbol{\tau}} = \underset{\boldsymbol{\tau} \in \mathcal{T}}{\operatorname{argmin}} \left(\hat{\boldsymbol{\mu}} - \frac{1}{S} \sum_{s=1}^{S} \hat{\boldsymbol{\mu}}^{(s)}(\boldsymbol{\tau}) \right)' \left(\hat{\boldsymbol{\mu}} - \frac{1}{S} \sum_{s=1}^{S} \hat{\boldsymbol{\mu}}^{(s)}(\boldsymbol{\tau}) \right), \qquad (3.3.8)$$

where $\mathcal{T}$ is the parameter space of $\boldsymbol{\tau}$. The vector $\hat{\boldsymbol{\mu}} = \left(\hat{\boldsymbol{\beta}}_{\setminus \hat{\boldsymbol{\phi}}}, \operatorname{var}(\hat{\phi})\right)$, where $\hat{\boldsymbol{\beta}}_{\setminus \hat{\boldsymbol{\phi}}}$ corresponds to the vector $\hat{\boldsymbol{\beta}}$ without $\hat{\boldsymbol{\phi}}$, and $\operatorname{var}(\hat{\phi}) = \frac{1}{k-1} \sum_{j=1}^{k} (\hat{\phi}_j - \bar{\hat{\boldsymbol{\phi}}})^2$ is the variance of $\hat{\boldsymbol{\phi}}$, with $\hat{\phi}_j$ being its j^{th} element and $\bar{\hat{\boldsymbol{\phi}}} = \frac{1}{k} \sum_{j=1}^{k} \hat{\phi}_j$. In the literature about indirect inference, the minimisation (3.3.8) is generally applied to $\hat{\boldsymbol{\beta}}$ directly[5], instead of a modified version of it, which in our case is $\hat{\boldsymbol{\mu}}$. However, in our setting matching the estimates for $\boldsymbol{\phi}$, based on true and simulated data, is inadvisable since this parameter vector bears information about the evolution of γ_t over time, not only about σ_γ, which we instead aim to estimate by indirect inference. The goal at this stage is to estimate only the time-constant parameters of the nonlinear model, not the evolution of its state variables over time, which we instead need in the filtering step (described in Section 3.3.2). This is the reason why the simulated paths, $Z_T^{(s)}(\boldsymbol{\tau})$, only need to depend on the values of $\boldsymbol{\tau}$ and need not be conditioned on the observed data, Z_T; the maximum likelihood estimates for $\boldsymbol{\phi}$ can be very different for distinct simulated paths, even if these paths are generated based on the same value of $\boldsymbol{\tau}$, and therefore of σ_γ. We therefore need a function of $\hat{\boldsymbol{\phi}}$ that is informative

[5] In this case Gourieroux et al. (1993) show that $\hat{\boldsymbol{\tau}}$ is a consistent estimator of $\boldsymbol{\tau}$, and that, for S fixed, $\sqrt{T}\,(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau})$ is asymptotically normally distributed with covariance matrix depending on the function relating $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, which is (as in our case) often unknown.

only about σ_γ; the statistics $\mathrm{var}(\hat{\phi})$ achieves this goal since $\hat{\phi}_j$ represents the value of the cubic spline estimate of γ_t at the j^{th} knot, for $j = 1, \ldots, k$, and $\mathrm{var}(\hat{\phi})$ is therefore a measure of the cubic spline estimate's spread. Intuitively, the larger σ_γ, the more volatile we would expect the cubic splines estimate of γ_t to be, and therefore the larger the value of $\mathrm{var}(\hat{\phi})$[6].

Finally, notice that $\dim(\boldsymbol{\mu}) \geq \dim(\boldsymbol{\tau})$, in order for $\boldsymbol{\tau}$ to be idenitifiable (Gourieroux et al., 1993). In our case the two vectors have the same dimension[7].

Algorithm 1 reports the detailed steps for the implementation of the indirect inference estimation of $\boldsymbol{\tau}$.

The stochastic specification of a time-varying parameter is more flexible than the deterministic one based on cubic splines, since it requires fewer assumptions on the parameter's evolution over time. With the cubic splines we indeed have to assume that the parameter is changing smoothly over time, and we need to choose the location of the knots a priori. The dynamic equation (random walk in our case) for the stochastic specification also needs to be chosen a priori, but (as will be shown in the Monte Carlo simulation study) is able to accommodate a wider variety of time-varying patterns. However, other non-random and more flexible time-varying specifications, than the one based on cubic splines, could potentially be used for the approximate model in the indirect inference approach. For instance, we also followed the idea of Delle Monache et al. (2016), who estimate time-varying parameters in state space models with the score-driven method of Creal et al. (2013) and Harvey

[6] Monte Carlo simulation results, that we do not report in this chapter, suggest that other measures, such as the variance of the cubic spline estimate of γ_t, can alternatively be employed in order to achieve the same goal. In principle, several other measures of the spline variability, also based on its second derivative, could be alternatively employed for the indirect inference estimation. Poirier (1973) provides the analytical expression for this second derivative, and it should be possible to start from them in order to build such a measure.

[7] We choose a random walk specification for γ_t, given in equation (3.3.5), because it is a rather flexible one, and it allows for structural changes in the correlation parameter. Generally, an AR(1) specification could be employed instead. However, in the latter case the inequality would not hold anymore, because the AR(1) specification implies that at least one additional parameter in the nonlinear model needs to be estimated.

(2013), by assuming that the time-varying correlation depends on its past values and past values of the score function (i.e., the scaled derivative of the log-likelihood with respect to the correlation). Nevertheless, Monte Carlo simulation results, which we report and discuss in Chapter 4, show that this method does not satisfactorily estimate the time-varying correlation. Moreover, this approach is computationally much more expensive than the one based on cubic splines. We therefore decided not to pursue it.

3.3.2 Rao-Blackwellised bootstrap filter

Once the static parameters of the nonlinear model have been estimated, we can proceed with the filtering step in order to estimate ρ_t, together with the other state variables of the model. The nonlinearity of the state space model remains after the estimation of $\boldsymbol{\tau}$, which means that alternative filters than the Kalman filter need to be used. As already anticipated in the Introduction, we employ the bootstrap filter of Gordon et al. (1993), which is a type of particle filter.

Before proceeding with the explanation of the method, we point out that the bootstrap filter is always applied to the nonlinear model (3.3.6), evaluated at the indirect inference estimate, $\hat{\boldsymbol{\tau}}$, even when we do not make this explicit in what follows. We therefore indicate with $\hat{\boldsymbol{T}}^*$, $\hat{\boldsymbol{H}}$, and $\hat{\boldsymbol{Q}}_t^*$ the indirect inference estimates for the corresponding matrices of model (3.3.6). Notice that $\hat{\boldsymbol{Q}}_t$ (from model (3.3.1)) is equivalent to $\hat{\boldsymbol{Q}}_t^*$ without its last row and column.

The idea of bootstrap filtering is to simulate, at time t and unconditionally on the data, $\boldsymbol{z}_t$, M values for each state variable of model (3.3.6), from the density distribution implied by their transition equation: draw $\boldsymbol{\alpha}_t^{(m)*} \sim N\left(\hat{\boldsymbol{T}}^*\boldsymbol{\alpha}_{t-1}^{(m)*}, \boldsymbol{R}^*\hat{\boldsymbol{Q}}_t^*\boldsymbol{R}^{*'}\right)$, for $m = 1, \ldots, M$. These M generated values are called particles. Of course, since the particles are simulated unconditionally on the data, some of them will be far from the true values of the state variables, and some others will be close. To make sure that the latter happens, M should be large. In order to understand which of the particles are closer to the true values, we compute the likelihood for each one of them: $p\left(\boldsymbol{z}_t|\boldsymbol{\alpha}_t^{(m)*}, \hat{\boldsymbol{\tau}}\right)^{(m)}$,

Algorithm 1 INDIRECT INFERENCE ESTIMATION OF $\boldsymbol{\tau}$

```
Estimate the cubic splines model on the observed data, $Z_T$, by Kalman filter and maximum likelihood. Store the maximum likelihood estimate, $\hat{\boldsymbol{\beta}}$.
Generate standard normal errors $\tilde{\boldsymbol{\varepsilon}}_t^{(s)} \sim N(\mathbf{0}, \boldsymbol{I}_q)$, and $\tilde{\boldsymbol{\eta}}_t^{(s)*} \sim N(\mathbf{0}, \boldsymbol{I}_{r+1})$, for $t = 1, \ldots, T$ and $s = 1, \ldots, S$. These simulated errors are used for *all* values of $\boldsymbol{\tau}$ in minimisation (3.3.8), which means that they have to be stored *before* performing the minimisation (otherwise the objective function used in equation (3.3.8) is not continuous with respect to $\boldsymbol{\tau}$).
**for** $\boldsymbol{\tau} \in \mathcal{T}$ **do**
    **for** $s \in \{1, ..., S\}$ **do**
        Set initial values $\boldsymbol{\alpha}_0^{(s)*} = \mathbf{0}$. Then generate new series $Z_T^{(s)}$ according to the non-linear model (3.3.6). This is done as follows:
        **for** $t \in \{1, ..., T\}$ **do**
            Let $\boldsymbol{C}_H$ be the Cholesky decomposition of $\boldsymbol{H}$. Get $\boldsymbol{\varepsilon}_t^{(s)} = \boldsymbol{C}_H \tilde{\boldsymbol{\varepsilon}}_t^{(s)}$.
            Get $\eta_{\gamma,t}^{(s)} = \sigma_\gamma \tilde{\eta}_{\gamma,t}^{(s)}$, and then $\gamma_t^{(s)} = \gamma_{t-1}^{(s)} + \eta_{\gamma,t}^{(s)}$.
            Evaluate $\boldsymbol{Q}_t^{(s)*}$ at $\gamma_t^{(s)}$. Let $\boldsymbol{U}_{Q,t}^{(s)}$ be the Cholesky decomposition of $\boldsymbol{Q}_t^{(s)*}$ without its last row and column (which is $\boldsymbol{Q}_t^{(s)}$). Get $\boldsymbol{\eta}_t^{(s)} = \boldsymbol{U}_{Q,t}^{(s)} \tilde{\boldsymbol{\eta}}_t^{(s)}$ (notice that $\tilde{\boldsymbol{\eta}}_t^{(s)}$ is equal to $\tilde{\boldsymbol{\eta}}_t^{(s)*}$ without its last element, which is $\tilde{\eta}_{\gamma,t}^{(s)}$).
            Get $\boldsymbol{\alpha}_t^{(s)} = \boldsymbol{T}\boldsymbol{\alpha}_{t-1}^{(s)} + \boldsymbol{R}\boldsymbol{\eta}_t^{(s)}$.
            Get $\boldsymbol{z}_t^{(s)} = \boldsymbol{Z}\boldsymbol{\alpha}_t^{(s)} + \boldsymbol{M}\boldsymbol{\varepsilon}_t^{(s)}$.
        **end for**
        Let $Z_T^{(s)} = \{\boldsymbol{z}_1^{(s)}, \ldots, \boldsymbol{z}_T^{(s)}\}$. Estimate the cubic splines model on the simulated data, $Z_T^{(s)}$, by Kalman filter and maximum likelihood. Store the maximum likelihood estimate, $\hat{\boldsymbol{\beta}}^{(s)}$, and respective $\hat{\boldsymbol{\mu}}^{(s)}$.
    **end for**
    Store $\left(\hat{\boldsymbol{\mu}} - \frac{1}{S}\sum_{s=1}^S \hat{\boldsymbol{\mu}}^{(s)}\right)' \left(\hat{\boldsymbol{\mu}} - \frac{1}{S}\sum_{s=1}^S \hat{\boldsymbol{\mu}}^{(s)}\right)$.
**end for**
Find the value of $\boldsymbol{\tau}$ which minimises $\left(\hat{\boldsymbol{\mu}} - \frac{1}{S}\sum_{s=1}^S \hat{\boldsymbol{\mu}}^{(s)}\right)' \left(\hat{\boldsymbol{\mu}} - \frac{1}{S}\sum_{s=1}^S \hat{\boldsymbol{\mu}}^{(s)}\right)$. This can be done for a grid of values of $\boldsymbol{\tau}$, or, more appropriately, with a gradient/Hessian-based optimisation algorithm that finds the solution numerically.
Notice that, for notation simplicity, in the algorithm we omitted the dependence on $\boldsymbol{\tau}$ of all simulated series and respective maximum likelihood estimates. However, it should be kept in mind that all elements with an $(s)$-superscript depend on $\boldsymbol{\tau}$.
```

for $m = 1, \ldots, M$. The larger the likelihood, the more likely the corresponding particle is close to the true value of $\boldsymbol{\alpha}_t^*$. We then build M weights that are proportional to the respective likelihoods, $\tilde{w}_t^{(m)} = w_{t-1}^{(m)} p\left(\boldsymbol{z}_t | \boldsymbol{\alpha}_t^{(m)*}, \hat{\boldsymbol{\tau}}\right)^{(m)}$, and we standardise them: $w_t^{(m)} = \frac{\tilde{w}_t^{(m)}}{\sum_{m=1}^{M} \tilde{w}_t^{(m)}}$, for $m = 1, \ldots, M$. Finally, we resample with replacement M particles with probabilities corresponding to the normalised weights $\left\{w_t^{(1)}, \ldots, w_t^{(M)}\right\}$, in order to make sure that we keep the particles that yield a larger likelihood. The bootstrap filter estimate for $\boldsymbol{\alpha}_t^*$ at time t is equal to $\frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\alpha}_t^{(m)*}$, where $\boldsymbol{\alpha}_t^{(m)*}$, for $m = 1, \ldots, M$, are the resampled particles. These resampled particles are then used at time $t+1$ in the expected value of the normal distribution from which new M vectors $\boldsymbol{\alpha}_{t+1}^{(m)*}$ are generated. The entire procedure then repeats again until $t = T$.

However, we do not stop at standard bootstrap filtering because when the state vector is large, which is the case for the extended Dutch labour force model, then the bootstrap filter can become computationally costly. We therefore notice that if γ_t is known at time t, then model (3.3.6) becomes linear again, as it is only γ_t that triggers the nonlinearity of the model. Therefore, if we could condition on γ_t at time t, then the other state variables, $\boldsymbol{\alpha}_t$, could be predicted by standard Kalman filtering, which is a very efficient estimation method. This conditioning can be achieved by Rao-Blackwellising the bootstrap filter, as proposed by Chen and Liu (2000). Namely, at time t we generate M particles only for γ_t: $\gamma_t^{(m)} \sim N\left(\gamma_{t-1}^{(m)}, \hat{\sigma}_\gamma\right)$, for $m = 1, \ldots, M$. Then the one-step ahead forecasts of the other state variables, $\boldsymbol{a}_{t+1}^{(m)}$, together with their variances, $\boldsymbol{P}_{t+1}^{(m)}$, can be obtained by running the prediction step of the Kalman filter recursions (3.3.4), applied to the linear model (3.3.1) evaluated at $\hat{\boldsymbol{\tau}}$ and with γ_t replaced by $\gamma_t^{(m)}$.

At time $t+1$, for each set of particles $\left\{\gamma_t^{(m)}, \boldsymbol{a}_{t+1}^{(m)}, \boldsymbol{P}_{t+1}^{(m)}\right\}$, it is then possible to obtain the prediction error, $\boldsymbol{v}_{t+1}^{(m)}$, together with its covariance matrix, $\boldsymbol{F}_{t+1}^{(m)}$, for $m = 1, \ldots, M$. These two elements can be used in order to evaluate the following likelihood:

$$p\left(\boldsymbol{z}_{t+1}|\gamma_t^{(m)}, \boldsymbol{a}_{t+1}^{(m)}, \boldsymbol{P}_{t+1}^{(m)}, \hat{\boldsymbol{\tau}}\right)^{(m)} = \\ \exp\left(-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left(\det \boldsymbol{F}_{t+1}^{(m)}\right) - \frac{1}{2}\boldsymbol{v}_{t+1}^{(m)\prime}\left(\boldsymbol{F}_{t+1}^{(m)}\right)^{-1}\boldsymbol{v}_{t+1}^{(m)}\right), \tag{3.3.9}$$

for $m = 1, \ldots, M$. The M likelihood values can in turn be employed to build the same standardised weights used in the standard bootstrap filter, which allow us to resample with replacement the sets of particles $\left\{\gamma_t^{(m)}, \boldsymbol{a}_{t+1}^{(m)}, \boldsymbol{P}_{t+1}^{(m)}\right\}$, for $m = 1, \ldots, M$, that yield larger likelihood values. The resampled particles can then be used in order to repeat the procedure again for the next period in time. Notice that we also need to resample $\boldsymbol{a}_{t+1}^{(m)}$ and $\boldsymbol{P}_{t+1}^{(m)}$ because they need to be used as inputs in the prediction step of the Kalman filter, in order to obtain $\boldsymbol{a}_{t+2}^{(m)}$ and $\boldsymbol{P}_{t+2}^{(m)}$, for $m = 1, \ldots, M$. The Rao-Blackwellised bootstrap filter (RBBF) estimate of γ_t at time t, which we indicate with $\hat{\gamma}_{\text{RBBF},t}$, is again obtained by taking the average of the M resampled particles $\gamma_t^{(m)}$. Algorithm 2 outlines the Rao-Blackwellised bootstrap filter estimation of ρ_t in detail.

Some remarks are now in place. First, the likelihood (3.3.9) for $\boldsymbol{z}_t$ at time t is conditioned on $\gamma_{t-1}^{(m)}$, i.e., on values for the time-varying correlation at the previous point in time. This is due to the Kalman filter recursions (3.3.4), where the prediction errors, $\boldsymbol{v}_t$, and respective covariance matrix, $\boldsymbol{F}_t$, depend on $\boldsymbol{Q}_{t-1}$, which contains γ_{t-1}, not $\boldsymbol{Q}_t$. Therefore, it is only possible to obtain the RBBF estimate of γ_t for $t = 1, \ldots, T-1$, and not for the last point in time, T. This is a consequence of the Rao-Blackwellisation as this issue does not arise in standard bootstrap filtering. In the latter case all state variables are generated simultaneously, whereas in the former only *predictions* of $\boldsymbol{\alpha}_{t+1}$ can be obtained based on sampled values for γ_t.

Second, the final estimates for the state vector $\boldsymbol{\alpha}_t$, can be obtained by running the standard Kalman filter recursions (3.3.4) applied to the linear model evaluated at $\hat{\boldsymbol{\tau}}$ and with γ_t replaced by $\hat{\gamma}_{t,\text{RBBF}}$, for $t = 1, \ldots, T-1$ (notice that the estimate for $\boldsymbol{\alpha}_t$ at time T is also unavailable). Although we do not cover it in this chapter, we would like to point out that a prediction for γ_T could be obtained, for instance, by taking the *unweighted* (i.e., not depending on likelihood values) average of the M particles $\gamma_T^{(m)}$, which in turn could be employed to predict $\boldsymbol{\alpha}_T$.

Third, a more efficient particle filter could be obtained by generating particles for γ_t, conditionally on the observed data, $\boldsymbol{z}_t$. This can be done by, for instance, using a sequential importance sampling approach, instead of the bootstrap filter. However, as pointed out in the Introduction, the type of nonlinearity we are dealing with challenges the quest for an importance density (i.e., a linear model that approximates the nonlinear one, in a more sophisticated way than the cubic splines model), which is needed in order to implement the above-mentioned approach (see Durbin and Koopman (2012, Chapter 12) and Creal (2012) for details about this and other methods).

Fourth, the resampling step of the algorithm is sometimes necessary in order to avoid the degeneracy of the particle filter over time (see again Durbin and Koopman (2012, Chapter 12) and Creal (2012) for an in-depth discussion of this problem). More sophisticated resampling methods, than the generalised one discussed above, can be employed. Li et al. (2015) provide an extensive review of all existing resampling methods for particle filtering. The ones that are shown to yield the particle filters with lowest Monte Carlo variation, are the stratified resampling of Kitagawa (1996) and the residual resampling of Liu and Chen (1998). We employ the former. However, before performing stratified resampling, we include an additional step. As mentioned in the Introduction, the likelihood function for $\boldsymbol{z}_t$, given in equation (3.3.9), is not that sensitive to different values of $\gamma_{t-1}^{(m)}$: in other words, the M particles $\gamma_{t-1}^{(m)}$ yield different values for the likelihood, but these differences are not large. Therefore, the corresponding normalised weights are similar to each other and any kind of resampling method will tend to select most of the particles, yielding a final RBBF estimate for γ_t that looks rather constant over time. In order to avoid this problem we first increase the differences among the normalised weights, which in turn increases the probability of resampling those particles that yield slightly larger likelihood values. We do so by following

Algorithm 2 Rao-Blackwellised bootstrap filter estimation of ρ_t

```
Initialise the filter at $t = 0$ (before the first observed period) with $\tilde{w}_0^{(m)} = 1$ (which implies $w_0^{(m)} = 1/M$), for $m = 1, \dots, M$. First sample $\gamma_0^{(m)} \sim N\left(\hat{\gamma}, \hat{\sigma}_\gamma^2\right)$, where $\hat{\gamma}$ is the maximum likelihood estimate of $\gamma$ from the model where this parameter is treated as static. Then draw $\boldsymbol{a}_1 \sim N\left(\boldsymbol{0}, \hat{\boldsymbol{Q}}_0^{(m)}\right)$, and set $\boldsymbol{P}_1^{(m)} = \hat{\boldsymbol{Q}}_0^{(m)}$, for $m = 1, \dots, M$. In this case $\boldsymbol{a}_t$ is the one-step ahead prediction of the Kalman filter for $\boldsymbol{\alpha}_t$, with $\boldsymbol{P}_t$ being its predicted covariance matrix, and $\hat{\boldsymbol{Q}}_0^{(m)}$ is evaluated at $\gamma_0^{(m)}$. We need initialisations for both $\boldsymbol{a}_t$ and $\boldsymbol{P}_t$ in order to implement the Kalman filter.
for $t \in \{1, ..., T\}$ do
    Use $\left\{\gamma_{t-1}^{(m)}, \boldsymbol{a}_t^{(m)}, \boldsymbol{P}_t^{(m)}\right\}$ in order to run the prediction step of the Kalman filter applied to the linear model evaluated at $\hat{\boldsymbol{\tau}}$, which yields $\boldsymbol{v}_t^{(m)}$ and $\boldsymbol{F}_t^{(m)}$. Then compute
        $p\left(\boldsymbol{z}_t | \gamma_{t-1}^{(m)}, \boldsymbol{a}_t^{(m)}, \boldsymbol{P}_t^{(m)}, \hat{\boldsymbol{\tau}}\right)^{(m)} = \exp\left(-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\left(\det \boldsymbol{F}_t^{(m)}\right) - \frac{1}{2}\boldsymbol{v}_t^{(m)\prime}\left(\boldsymbol{F}_t^{(m)}\right)^{-1}\boldsymbol{v}_t^{(m)}\right),$
    for $m = 1, \dots, M$.
    Compute the weights $\tilde{w}_t^{(m)} = w_{t-1}^{(m)} p\left(\boldsymbol{z}_t | \gamma_{t-1}^{(m)}, \boldsymbol{a}_t^{(m)}, \boldsymbol{P}_t^{(m)}, \hat{\boldsymbol{\tau}}\right)^{(m)}$ for $m = 1, \dots, M$. Then obtain the normalised weights $w_t^{(m)} = \frac{\tilde{w}_t^{(m)}}{\sum_{m=1}^M \tilde{w}_t^{(m)}}$, for $m = 1, \dots, M$.
    Resample $M$ particles $\left\{\gamma_{t-1}^{(m)}, \boldsymbol{a}_t^{(m)}, \boldsymbol{P}_t^{(m)}\right\}$, with $m = 1, \dots, M$, with replacement, based on the modified stratified resampling technique (such modification is discussed at the end of Section 3.3.2).
    Reset $w_t^{(m)} = 1/M$ for $m = 1, \dots, M$.
    Compute $\hat{\gamma}_{\text{RBBF},t-1} = \frac{1}{M}\sum_{m=1}^M \gamma_{t-1}^{(m)}$, which is the Rao-Blackwellised bootstrap filter estimate for $\gamma_{t-1}$. To get the corresponding estimate for $\rho_{t-1}$ we take the hyperbolic tangent of $\hat{\gamma}_{\text{RBBF},t-1}$.
    Draw $\gamma_t^{(m)} \sim N\left(\gamma_{t-1}^{(m)}, \hat{\sigma}_\gamma^2\right)$ and use it with the resampled particles $\left\{\boldsymbol{a}_t^{(m)}, \boldsymbol{P}_t^{(m)}\right\}$ in order to run the prediction step of the Kalman filter applied to the linear model evaluated at $\hat{\boldsymbol{\tau}}$ and with $\gamma_t$ replaced by $\gamma_t^{(m)}$, which yields $\left\{\boldsymbol{a}_{t+1}^{(m)}, \boldsymbol{P}_{t+1}^{(m)}\right\}$, for $m = 1, \dots, M$.
end for
The final estimates for $\boldsymbol{\alpha}_t$ can be obtained by running the usual Kalman filter applied to the linear model evaluated at $\hat{\boldsymbol{\tau}}$ and with $\gamma_t$ replaced by $\hat{\gamma}_{t,\text{RBBF}}$, for $t = 1, \dots, T-1$.
```

Chen et al. (2001), who take functions of the normalised weights, $\left[w_t^{(m)}\right]^{p_t}$

for $m = 1, \ldots, M$, where $p_t \geq 0$ and depends on the coefficient of variation, which takes the expression

$$CV_t = \left[\frac{1}{M} \sum_{m=1}^{M} \left(M w_t^{(m)} - 1 \right)^2 \right]^{0.5},$$

and is a measure of weight instability. It varies between 0 and $\sqrt{M-1}$. If all weights are equal, then CV_t is equal to its lower bound (Creal, 2012). If the goal is to give more presence to the particles with larger weights, which is our case, then $p_t > 1$ if the coefficient of variation is low, otherwise $p_t < 1$. We use $p_t = \ln\left(\frac{\sqrt{M-1}}{CV_t}\right)$ in order to achieve this goal. Once the weights have been transformed according to this function, they have to be re-standardised before performing stratified resampling.

Finally, for simplicity we sometimes state, throughout the chapter, that the RBBF is employed to estimate the entire state vector of the nonlinear model. However, it should now be clear that it is really only ρ_t that is estimated by the RBBF, and the remaining part of the state vector is estimated by Kalman filtering with ρ_t replaced by its RBBF estimate. The above-mentioned simplification therefore only helps us to point out that it is the RBBF that solves the nonlinearity of the model, and that therefore allows us to estimate the state vector.

3.4 Monte Carlo simulation study

We conduct a Monte Carlo simulation study to assess the performance of the two estimation methods proposed in Section 3.3: the cubic splines method, which already is, per se, an estimation method of the time-varying state correlation, and the method based on the combination of indirect inference and Rao-Blackwellised bootstrap filtering. The performance of the methods is evaluated in several ways. First, we want to assess whether the indirect inference approach appropriately estimates the static parameters of the nonlinear model. Secondly, we check if the proposed methods are able to estimate the true time-varying relationships, and, in that case, to what level of accuracy. Finally, since in the empirical application our variable of interest, Dutch unemployment, is assumed to be unknown and therefore enters the model as an unobserved component, we want to understand to what extent the accuracy of the estimation of the state variables of interest changes if we take the time-varying relationships between the observed series into account.

For the sake of computational time, in the Monte Carlo simulation study we consider the following bivariate local level model, which is more simple than the Dutch Labour Force model:

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{Z}\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N\left(\mathbf{0}, \boldsymbol{H}\right) \\ \boldsymbol{\alpha}_{t+1} &= \boldsymbol{T}\boldsymbol{\alpha}_t + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N\left(\mathbf{0}, \boldsymbol{Q}_t\right), \quad t = 1, \ldots, T, \end{aligned}$$

where $\boldsymbol{z}_t = (y_t, x_t)'$, $\boldsymbol{\alpha}_t = (L_{y,t}, L_{x,t})'$, $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are all 2×1 vectors, $\boldsymbol{Z} = \boldsymbol{T} = \boldsymbol{I}_2$, $\boldsymbol{H} = \text{diag}(\sigma^2_{\varepsilon,y}, \sigma^2_{\varepsilon,x})$, with $\sigma_{\varepsilon,y} = \sigma_{\varepsilon,x} = 1$, and $\boldsymbol{Q}_t = \begin{bmatrix} \sigma^2_{\eta,y} & \rho_t\sigma_{\eta,y}\sigma_{\eta,x} \\ \rho_t\sigma_{\eta,y}\sigma_{\eta,x} & \sigma^2_{\eta,x} \end{bmatrix}$, with $\sigma_{\eta,y} = \sigma_{\eta,x} = 1$ (we are also implicitly imposing $\boldsymbol{M} = \boldsymbol{R} = \boldsymbol{I}_2$).

In the model above, $\boldsymbol{z}_t$ represents the observed vector, $\boldsymbol{\alpha}_t$ the vector of state variables, $\boldsymbol{\varepsilon}_t$ the vector of innovations in the measurement equation, and $\boldsymbol{\eta}_t$ the vector of innovations in the transition equation. We assume that the state variable of the first observable series, $L_{y,t}$, is the unobserved component of interest.

We consider the following data generating processes (DGPs) for the time-varying parameter, ρ_t, which are partly inspired by Creal et al. (2011) and Delle Monache et al. (2016):

1. Constant: $\rho_t = 0.9$
2. Sine: $\rho_t = 0.5 + 0.4\cos(2\pi t/(T/3))$
3. Fast sine: $\rho_t = 0.5 + 0.4\cos(2\pi t/(T/6))$
4. Step: $\rho_t = 0.9 - 0.5(t > T/2)$

5. Ramp: $\rho_t = 2/T \mod (t/(T/2))$

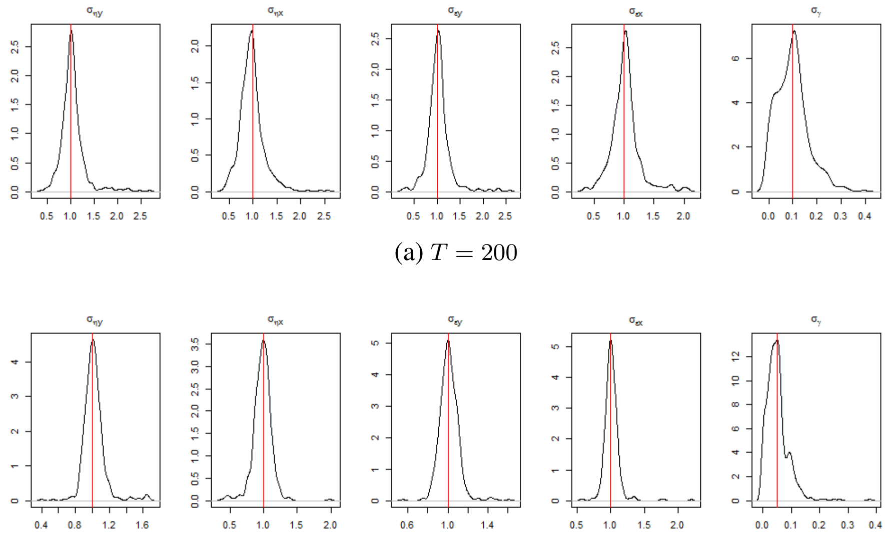
6. Random walk: $\rho_t = \tanh(\gamma_t)$, where $\gamma_t = \gamma_{t-1} + \eta_{\gamma,t}$, with $\eta_{\gamma,t} \sim N(0, \sigma_\gamma^2)$,

for $t = 1, \ldots, T$. We consider two different sample sizes: $T = 200$, which is close to the sample size of our empirical application, and $T = 500$. For the random walk specification of ρ_t, we set $\sigma_\gamma = 0.1$ when $T = 200$ and $\sigma_\gamma = 0.05$ when $T = 500$, to make sure that ρ_t does not become too volatile with a larger sample size (since the variance of a random walk process increases with time). An improvement in estimation performance is to be expected with a larger sample size. We run $n_{\text{sim}} = 500$ Monte Carlo simulations and always use $S = 3$ simulations for the indirect inference method, and $M = 5000$ particles for the RBBF. We use the BFGS algorithm to solve the minimisation problem, given in equation (3.3.8), that finds the indirect inference estimates.

We start the discussion of the Monte Carlo simulation results by looking at the performance of the indirect inference approach in estimating the static parameters of the nonlinear model. Figure 3.4.1 shows the density distributions of the indirect inference estimators of the elements of $\boldsymbol{\tau} = (\sigma_{\eta,y}, \sigma_{\eta,x}, \sigma_{\varepsilon,y}, \sigma_{\varepsilon,x}, \sigma_\gamma)'$, based on the Monte Carlo replicates, together with the true values of these parameters, when the DGP for ρ_t is a random walk. All distributions are centered around the true values, and their spread decreases with the sample size. For σ_γ, the distribution of its indirect inference estimator is skewed to the right and shows a bump around values of the parameter close to zero, which tends to be less pronounced, but does not disappear, with a larger sample size . This issue seems somewhat similar to the pile-up problem discussed, among others, by Shephard and Harvey (1990) and Stock and Watson (1998): if the scale parameter (in our case σ_γ) of a coefficient that varies stochastically over time by following a random walk specification, is small, then its maximum likelihood estimator has a point mass at zero, and this probability mass decreases with a larger sample size. The results shown in Figure 3.4.1 are obtained by using the true values of the parameters as initial values in the BFGS algorithm. In practice, a good starting value for σ_γ is difficult to find. Figure 3.A.1 therefore reports the same results when the initial value for σ_γ is twice as big as its true one. When $T = 200$, the algorithm gets stuck around the initial value half of the times, making the distribution of the indirect inference estimator of σ_γ look bimodal. This problem, however, does not occur any more when the sample size increases; in this case the distribution is very similar to the one shown in Figure 3.4.1. We therefore advise practitioners who deal with small sample sizes, to first evaluate the objective function that needs to be minimised in order to find the indirect inference estimates (from equation (3.3.8)), for a grid of values for σ_γ, in order to come up with a good starting value for this parameter (we do so ourselves in the empirical application of Section 3.5[8]). Figures 3.A.2-3.A.6 show the same results for the remaining five DGPs of ρ_t. All distributions, for all parameters except σ_γ, are again symmetrical and centered around the true values of the parameters, and their spread decreases with the sample size. In these cases, however, we do not know the true value for σ_γ, so we cannot assess whether its indirect inference estimator is centered around it. The shapes of the distributions for the latter estimator are similar to the ones observed for the random walk specification of ρ_t. The BFGS algorithm in these cases is initialised at $\sigma_\gamma = 0.1$ when $T = 200$ and $\sigma_\gamma = 0.05$ when $T = 500$.

Next we investigate the accuracy of our methods in estimating the time varying correlation and whether treating ρ_t as time-varying also improves the estimation of the state variable of interest. We indicate with $\hat{L}_{y,t}$ the Kalman filter estimator (i.e., the first element of $\boldsymbol{a}_{t|t}$ from the Kalman filter recursions (3.3.4)) of $L_{y,t}$, and with $\hat{\rho}_t$ the estimator of ρ_t (whether this estimator is based on the cubic splines or the RBBF method, will be clear in the discussion that follows). The Mean Squared Error (MSE) and the squared bias of $\hat{\rho}_t$ are com-

[8] The grid values are chosen arbitrarily. However, since ρ_t is bounded to be between -1 and 1, σ_γ cannot get too large values, and this restricts the dimension of its space (in practice).



(a) $T = 200$

(b) $T = 500$

Figure 3.4.1: Distribution of the indirect inference estimators of the static parameters of the nonlinear model, $\boldsymbol{\tau} = (\sigma_{\eta,y}, \sigma_{\eta,x}, \sigma_{\varepsilon,y}, \sigma_{\varepsilon,x}, \sigma_{\gamma})'$, based on the Monte Carlo replicates, when the DGP of ρ_t is a random walk; $S = 3$, $n_{\text{sim}} = 500$. The red lines represent the true values of the parameters.

puted as follows:

$$\text{MSE}\left(\hat{\rho}_t\right) = \frac{1}{T-1-d} \sum_{t=d+1}^{T-1} \left(\frac{1}{n_{\text{sim}}} \sum_{s=1}^{n_{\text{sim}}} \left(\hat{\rho}_{t,s} - \rho_{t,s}\right)^2 \right),$$

$$\text{bias}^2\left(\hat{\rho}_t\right) = \frac{1}{T-1-d} \sum_{t=d+1}^{T-1} \left(\frac{1}{n_{\text{sim}}} \sum_{s=1}^{n_{\text{sim}}} \left(\hat{\rho}_{t,s} - \rho_{t,s}\right) \right)^2 .$$

The same measures of fit are obtained also for $\hat{L}_{y,t}$ by substituting ρ_t by $L_{y,t}$ in the formulae above ($d = 0$ for the measures of $\hat{\rho}_t$, and $d = 1$ for the measures of $\hat{L}_{y,t}$, since $L_{y,t}$ is the only state variable of the target series which requires a diffuse initialisation). The MSE captures both the variance and the bias of the estimators, and therefore gives an indication (also) of their volatility. The squared bias, instead, is supposed only to capture the difference between the estimated and the true values. We use the square of the bias because we deal with estimators of time-varying parameters/variables, and therefore averaging the biases over time, without taking their squared values, would provide a misleading measure of comparison among different methods. Table 3.4.1 reports these measures of fit, obtained with both the cubic splines and the RBBF methods, relative to the same measures observed while estimating the correlation as time-constant (i.e., by maximum likelihood). We refer to Table 3.A.1 for the absolute figures. We indicate with "ideal Rao-Blackwellised bootstrap filter" the setting where the static parameter vector $\boldsymbol{\tau}$ is not estimated by indirect inference, but it is treated as known by using its true values (which is only possible when the DGP of ρ_t is a random walk), hence the term "ideal". This should allow us to quantify the influence that the estimation of the static parameters by indirect inference has on the performance of the RBBF. The Tables show that, when $T = 200$, the MSE of $\hat{\rho}_t$ is almost always better when the correlation parameter is estimated as constant, since the other two methods are, because of their time-varying nature, obviously more volatile. The latter methods, however, tend to strongly beat the time-constant estimation of the correlation in terms of squared bias, indicating that they are able to capture its time-variation (when this is present). The cubic splines shows a better performance than the RBBF, due to its milder volatility, except when the true DGP for ρ_t is a fast sine; the RBBF seems therefore more suited than the cubic splines method in estimating rapidly-changing time variations in the correlation parameter. The ideal RBBF only shows a marginal improvement over the RBBF, suggesting that the estimation of the static parameters by indirect inference does not strongly influence the performance of the filter. All measures of fit for the two methods improve when the sample size increases, also with respect to estimating ρ_t as time-constant. When the true correlation is static, however, estimating it as such is preferred.

Since time-variation is the focal point of this chapter, it is also interesting to investigate how the two estimation methods perform over time. Figures 3.4.2 and 3.4.3 display, for each deterministic DGP of ρ_t, the 5%, 20%, 80% and 95% percentiles of the Monte Carlo simulation estimates (which we loosely

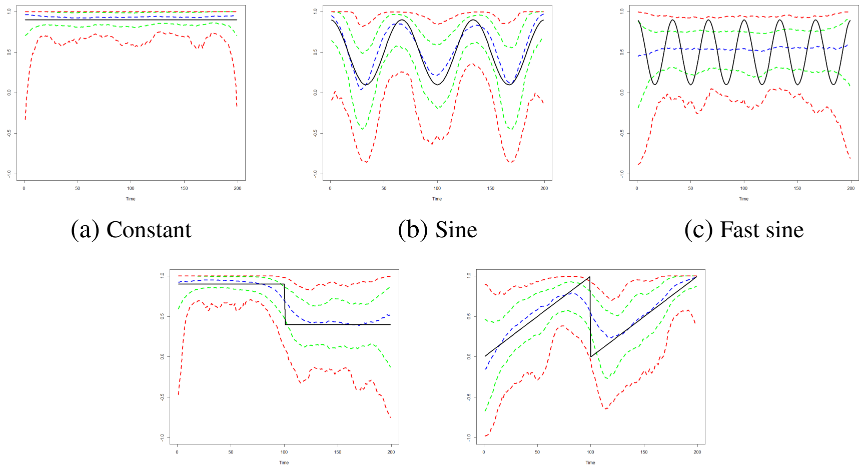
	$T = 200$						$T = 500$					
	0.9	Sine	Fast sine	Step	Ramp	Random walk	0.9	Sine	Fast sine	Step	Ramp	Random walk
	Cubic splines											
MSE($\hat{\rho}_t$)	25.769	1.688	2.194	1.269	1.345	0.920	8.114	0.615	1.364	0.454	0.643	0.574
bias^2($\hat{\rho}_t$)	45.540	0.086	1.016	0.069	0.178	0.505	29.112	0.123	1.003	0.062	0.165	0.001
MSE($\hat{L}_{y,t}$)	1.015	1.002	1.015	0.983	0.993	0.982	1.004	0.983	1.003	0.970	1.019	0.988
bias^2($\hat{L}_{y,t}$)	0.948	1.048	0.982	0.948	0.978	0.955	0.997	0.991	0.988	0.953	1.000	1.043
	Rao-Blackwellised bootstrap filter											
MSE($\hat{\rho}_t$)	28.118	1.835	1.716	1.565	1.775	1.269	15.317	1.504	1.207	1.048	1.457	1.074
bias^2($\hat{\rho}_t$)	47.041	0.623	0.845	1.813	0.442	1.191	97.757	0.370	0.878	0.228	0.337	0.003
MSE($\hat{L}_{y,t}$)	1.889	1.121	1.083	1.296	1.236	1.175	1.057	1.074	1.011	1.051	1.087	1.026
bias^2($\hat{L}_{y,t}$)	1.858	1.095	0.991	1.403	1.098	1.030	1.001	1.036	0.995	1.018	1.039	1.063
	Ideal Rao-Blackwellised bootstrap filter											
MSE($\hat{\rho}_t$)						1.183						1.118
bias^2($\hat{\rho}_t$)						1.070						0.002
MSE($\hat{L}_{y,t}$)						0.997						1.004
bias^2($\hat{L}_{y,t}$)						1.017						1.043

Table 3.4.1: Mean squared error and squared bias for the cubic splines and the Rao-Blackwellised bootstrap filter estimators of ρ_t, and the Kalman filter estimator of $L_{y,t}$, *relative* to the same measures of fit obtained while estimating ρ_t as time-constant. The second row lists the DGPs for ρ_t. "Ideal Rao-Blackwellised bootstrap filter" indicates that the static parameter vector $\boldsymbol{\tau}$ is treated as known by using its true values . $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.

call confidence bands), together with the median, obtained, respectively, with the cubic splines and the RBBF methods, when $T = 200$ (Figures 3.A.7 and 3.A.8 show the same results when $T = 500$). The results discussed above are confirmed by these Figures. Both methods are able to pick up the true time-variation of the correlation parameter, and their estimation performance improves with a larger sample size. The RBBF is therefore robust to misspecifications of the dynamic structure of ρ_t since none of the true correlations displayed in the Figures is a random walk. The confidence bands for the cubic splines are narrower, except towards the end of the sample, and less volatile than their RBBF's counterparts, but the fast sine DGP of ρ_t is estimated almost as constant by the cubic splines method, contrary to the RBBF. Notice also that the RBBF is slightly delayed in time with respect to the true time variation in ρ_t, which is to be expected from filters, since they are only based on past information, contrary to the time-constant estimation and the cubic splines (a particle smoother would not show this delay). Although the indirect inference method estimates the static parameters based on the entire sample, the RBBF can therefore be considered a real-time estimator. For this reason the time-constant and cubic splines methods are rewarded too much by the simulation study, with respect to the RBBF. A real-time comparison of the methods would tend to favour more the latter. In a simulation study, van den Brakel and Krieg (2016) showed that with a time constant correlation between survey and claimant counts data, it takes more than a year before the maximum likelihood estimate for the correlation picks up a change in the relation between the two series. However, for computational purposes, we only perform such a comparison in the empirical application of Section 3.5, since it is also of interest for the production of official statistics. The scope of the simulation study is just to understand to what extent the employed methods appropriately estimate, on average, the (time-varying) parameters.

Figures 3.A.9-3.A.12 report the absolute (i.e., not relative) MSE and squared bias of $\hat{\rho}_t$, over time, for all estimations methods (including the method that estimates the correlation as static). These pictures reveal information that was hidden in Tables 3.4.1 and 3.A.1, but already partly discovered from Figures 3.4.2, 3.4.3, 3.A.7 and 3.A.8: a better performance of the RBBF over the cubic splines, in terms of MSE, at the end of the sample, when $T = 200$. This is due to the large uncertainty that affects the cubic splines at the boundaries of the sample (we cannot conclude the same for the start of the sample because the performance of the RBBF there very much depends on the initial values for γ_t). In terms of squared bias and when $T = 500$, the two methods tend instead to perform similarly while approaching the end of the sample. What also appears from these Figures is that, when there are structural changes (i.e., when the true DGP of ρ_t is either a step or a ramp function) the RBBF is much worse in estimating the correlation *at* the change point in time, with respect to the other methods, due to its delayed behaviour. However, a real-time comparison would, also in this case, reward less the methods based on maximum likelihood.

Finally, the relative MSE and squared bias of $\hat{L}_{y,t}$ in Table 3.4.1 are generally around 1 for all DGPs of ρ_t and for both estimation methods, also when the sample size is large. This result indicates that estimating the correlation as time-varying, also when appropriate, instead of time-constant, does not have an impact on the estimation accuracy of the state variable of interest, probably



(a) Constant

(b) Sine

(c) Fast sine

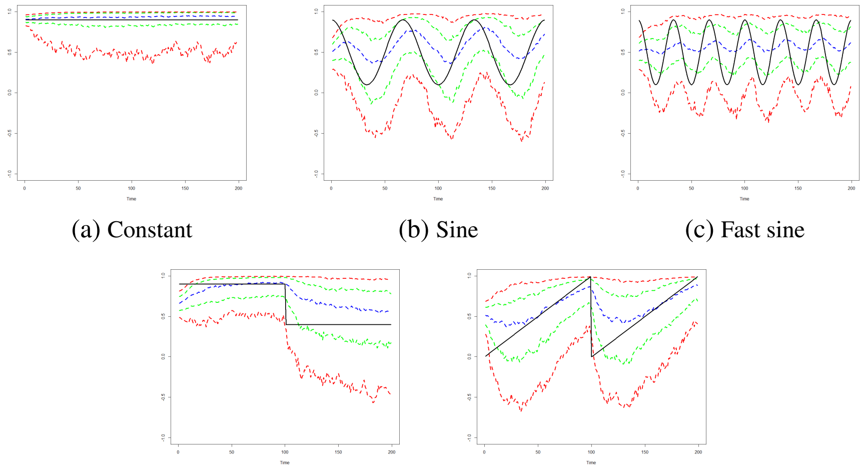
(d) Step

(e) Ramp

Figure 3.4.2: True process of ρ_t (black) together with the 95% (red), 80% (green) confidence bands, and the median (blue) of the simulation estimates for the cubic splines estimator of ρ_t. $T = 200$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.

due to the Kalman filter's high accuracy in estimating unobserved components, irrespectively of the parameters' estimates. Nonetheless, results from Section 3.5 suggest that in a real-time estimation these conclusions would be different. Figures 3.A.13-3.A.16 show the absolute (i.e., not relative) MSE and the squared bias of $\hat{L}_{y,t}$ over time, for all estimation methods. Except for a slightly worse performance, in terms of MSE, of the RBBF when $T = 200$, and a slightly better one of the cubic splines when $T = 500$, it is almost impossible to tell the difference, among methods, in the estimation accuracy of $L_{y,t}$. What appears evident from these Figures, however, is how the magnitude of the MSE tends to follow the time-varying pattern of ρ_t (for its deterministic DGPs): the MSE is lower when the magnitude of the correlation is larger, as the auxiliary series brings in this case more information about the state variable of interest.

In conclusion, the Monte Carlo simulation study shows that our proposed indirect inference method is able to correctly estimate the time-invariant pa-



(a) Constant

(b) Sine

(c) Fast sine

(d) Step

(e) Ramp

Figure 3.4.3: True process of ρ_t (black) together with the 95% (red), 80% (green) confidence bands, and the median (blue) of the simulation estimates for the Rao-Blackwellised bootstrap filter estimator of ρ_t. $T = 200$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.

rameters of the nonlinear model. Only the finite-sample distribution of the indirect inference estimator of σ_γ is not symmetrical, suggesting that it is not normal either. Nevertheless, since we do not carry out any inference on this parameter, we do not need normality of the estimator to hold[9]. Both the cubic splines and the RBBF approaches are suited to estimating a time-varying state correlation. The former method, on average, always beats the latter, in terms of estimation accuracy of ρ_t, due to the strong volatility that affects the RBBF, except when it comes to estimating correlations that change rapidly over time. Moreover, with small sample sizes, the RBBF yields more precise estimates for ρ_t towards the end of the sample, which is relevant when the focus is on

[9] Notice that providing theoretical properties for the indirect inference estimators is beyond the scope of this chapter. However, we point out that the knowledge of such asymptotic distributions would allow us to test for time-constancy in the correlation parameter, by testing whether $\sigma_\gamma = 0$.

exploring how the relationship between the observed series has evolved in recent times. The estimation accuracy of all methods (indirect inference, cubic splines and RBBF) improves with the sample size. Appropriately estimating the time-varying correlation as such, instead of as time-constant, does not affect the estimation accuracy of the state variables of interest. The RBBF is a real-time estimator of the correlation (i.e., it only exploits past information), whereas the time-constant and cubic splines methods estimate this parameter based on the entire sample. The simulation results obtained here therefore tend to reward the latter approaches too much. In real-time, the RBBF would yield an even better performance compared to the other two methods. Such a real-time comparison is not explored in the Monte Carlo simulation study, but is investigated in the upcoming Section.

3.5 Empirical application to the extended Dutch labour force model

In this Section we perform the in-sample estimation of the Dutch labour force model extended with the univariate series of claimant counts, described in Section 3.2. We model the time-varying correlation with our proposed methodology based on indirect inference and Rao-Blackwellised bootstrap filtering. As in the Monte Carlo simulation study, we compare the results so obtained to estimating the correlation parameter as time constant, and with the cubic splines method only. We also investigate the sensitivity of our results to the location of the cubic splines' knots. Specifically, we first consider the case where the (five, as suggested by the BIC) knots correspond to the quartiles of the sample. Then, we examine the case when February 2015 is one of the knots (since we know that a potential change in the correlation happened on this date), and the remaining (four) knots are chosen accordingly in order to keep an approximately equal distance between each pair of knots[10]. Finally, we compare the performance of the methods when

[10] The quartiles of the sample corresponds to the 0^{th}, 25^{th}, 50^{th}, 75^{th} and 100^{th} percentiles of the sample. When February 2015 is one of the knots, the knots are located at the 0^{th}, 28^{th}, 55^{th}, 69^{th} and 100^{th} percentiles of the sample.

the estimation of the model is conducted in real-time, and not only on the entire sample. We use monthly data from January 2004 until March 2020 ($T = 195$) for both the GREG estimates and the claimant counts, $S = 5$ simulations for the indirect inference method, and $M = 5000$ particles for the RBBF.

Figure 3.5.1 displays the estimated correlation, ρ_t, and its unbounded counterpart, γ_t, with all three methods, when the knots are located at the quartiles of the sample. The time-constant estimate for ρ_t is positive and very large, but it hides the decrease of the correlation parameter that occurred in the middle of the sample. Indeed, both the cubic splines and the RBBF estimates of ρ_t indicate that the two types of observed series started deviating from each other in 2010, that is, around two years after the financial crisis of 2008. Recession periods can induce long-term unemployment, and in the Netherlands unemployment benefits cannot be claimed for more than three years and two months, whether someone is unemployed or not, at the end of this period. However, not everyone is entitled to receive benefits for the maximum amount of time. The long-term unemployment caused by the economic crisis of 2008 is therefore not entirely picked up by the claimant counts series, which hence starts deviating from the GREG estimates around two years after the burst of the crisis. The magnitude of the correlation increased again in 2013, and stabilized around its ante-crisis levels from 2017 until the most recent times. Remember from the results of the Monte Carlo simulation study, that it is the RBBF estimate that provides the most reliable information on the behaviour of the parameter towards the end of the sample, which is indeed when the RBBF and the cubic splines estimates deviate the most from each other. We also notice how the RBBF, but not the cubic splines, captures a drop in the correlation after the legislative change of February 2015. New legislation often applies to new claims of unemployment benefits, which explains why the correlation does not drop immediately in February 2015, but shortly after. This decrease is, however, not as deep and prolonged as the one caused by the financial crisis of 2008.

Notably, Figure 3.5.1 further shows how the RBBF method yields a much more volatile estimate for ρ_t, than the cubic splines, which partly reflects estimation error. We already concluded from the simulation study of Section

		Knots at quartiles		Knots include February 2015	
	Maximum likelihood		Indirect inference	Maximum likelihood	Indirect inference
	Constant	Cubic splines		Cubic splines	
$\hat{\sigma}_{R,y}$	2719.916	3011.516	2996.161	2953.373	2971.015
$\hat{\sigma}_{\omega,y}$	0.016	0.028	0.02	0.044	0.044
$\hat{\sigma}_{\lambda}$	3778.169	3739.127	3831.573	3695.226	3690.149
$\hat{\sigma}_{\nu_1}$	1.335	1.322	1.261	1.328	1.331
$\hat{\sigma}_{\nu_2}$	1.309	1.304	1.260	1.3	1.272
$\hat{\sigma}_{\nu_3}$	1.199	1.192	1.106	1.186	1.176
$\hat{\sigma}_{\nu_4}$	1.234	1.234	1.173	1.235	1.229
$\hat{\sigma}_{\nu_5}$	1.219	1.214	1.179	1.22	1.22
$\hat{\delta}$	0.378	0.38	0.327	0.378	0.375
$\hat{\sigma}_{R,CC}$	3422.074	3453.365	3371.296	3485.617	3282.554
$\hat{\sigma}_{\omega,CC}$	0.019	0.021	0.02	0.01	0.01
$\hat{\sigma}_{\varepsilon,CC}$	1499.447	1481.391	1479.833	1462.325	1466.448
$\hat{\rho}$	0.851				
$\hat{\phi}_1$		0.367		-0.629	
$\hat{\phi}_2$		2.484		2.39	
$\hat{\phi}_3$		0.127		-0.098	
$\hat{\phi}_4$		1.697		1.386	
$\hat{\phi}_5$		0.333		0.269	
$\hat{\sigma}_{\gamma}$			0.081		0.127
Log-likelihood	-12549.42	-12546.34	-12476.98	-12547.15	-12475.59

Table 3.5.1: The columns named "Maximum likelihood" report the maximum likelihood estimates of the static parameters of the Dutch labour force model extended with the auxiliary series of claimant counts (described in Section 3.2), when the correlation parameter is estimated as time constant and with the cubic splines method. The remaining columns show the indirect inference estimates of the static parameters (including σ_γ). "Knots at quartiles" means that the knots for the cubic splines approach correspond to the quartiles of the sample; they are otherwise approximately equally distant by fixing February 2015 as a knot. The log-likelihood values are obtained by evaluating the Kalman filter recursions (3.3.4) at the corresponding estimates for ρ_t and the static parameters.

3.4, that the volatility of the RBBF estimator represents its Achilles' heel. Although we do not implement them in the chapter, we here give two suggestions that could potentially correct for this problem. First, a bootstrap smoother (instead of a filter) should yield, as its name gives away, a smoother estimate for ρ_t. Second, we explain in Section 3.3.2 how we modify the resampling step of the RBBF's algorithm, in order to keep particles that yield slightly larger likelihood values. We mentioned that without this modification, the filtered estimate for the correlation would almost be constant over time. We arbitrarily decided to take the $p_t = \ln\left(\frac{\sqrt{M-1}}{CV_t}\right)$ power of the normalised weights in order to achieve our goal. However, this function may increase the differences among weights too much and therefore make the RBBF too volatile. Many other functions of the normalised weights can be applied instead and potentially diminish the volatility of the filter.

Figure 3.B.1 compares the cubic splines and RBBF estimates for ρ_t that have been discussed above, with the ones obtained for the second choice of knots' location (which uses February 2015 as one of the knots). The estimates are not very sensitive to the location of the knots, if not for a slightly larger volatility of the cubic splines estimate with the second choice of knots' location, which is inherited also by the RBBF estimate.

Table 3.5.1 reports the estimates of the static parameters for each considered model. The indirect inference estimate for σ_γ is indeed larger when February 2015 is part of the knots. The table also displays the log-likelihood values that are obtained by evaluating the Kalman filter recursions (3.3.4), and subsequent likelihood function (3.3.2), at the corresponding estimates for ρ_t and the static parameters. Although we cannot perform a formal test on these log-likelihood values, because we are comparing non-nested models, we can still conclude that the estimation based on indirect inference and RBBF always yields a better fit to the data.

We employ the BFGS and a conjugate gradient optimisation algorithm when we estimate the static parameters, respectively, by maximum likelihood and by indirect inference. The complexity of the Dutch labour force model hampers its estimation by indirect inference with the BFGS algorithm[11]. When estimating the static parameters by maximum likelihood, we use their estimates from the labour force model obtained in van den Brakel and Krieg (2015), as initial values. The initial values of the static parameters in the indirect inference estimation are instead equal to the corresponding maximum likelihood estimates obtained with the cubic splines model.

[11] The algorithms hardly moves away from the initial values of the static parameters.

Figure 3.5.2 plots the Kalman filter estimates and respective estimated variances (i.e., the corresponding elements of $\boldsymbol{a}_{t|t}$ and $\boldsymbol{P}_{t|t}$ in the Kalman filter recursions (3.3.4)) of the state variables of interest $\theta_{y,t}$, $L_{y,t}$ and $R_{y,t}$, in the extended Dutch labour force model. The results refer to the setting where the knots correspond to the quartiles of the sample. The point estimates of the state variables are slightly sensitive to the estimation method of ρ_t, especially in the middle of the sample where the difference between time-constant and time-varying estimates for ρ_t is largest. The estimated variances are much more dependent on the magnitude of the correlation's estimates: they are larger when the estimated correlation shrinks, i.e., when the claimant counts bring less information about the Dutch unemployment. This is especially evident for $R_{y,t}$, than the other two state variables, since the slopes of the series' trends are directly related to each other via ρ_t. The estimated variances, when the correlation is estimated as time-constant, are generally very low (given the large value of the correlation's estimate). This may suggest that treating ρ_t as time-constant improves the estimation accuracy of state variables of interest. However, these variances are, in this case, wrongly estimated if the true process for ρ_t is time-varying. The variance estimates obtained with the cubic splines and the RBBF methods, although larger in the middle of the sample, are therefore more realistic as they reflect the economic uncertainty of that period. We again notice a better performance of the RBBF, compared to the cubic splines method, in the estimation accuracy of the state variables towards the end of the sample. Notice that these estimated variances do not reflect the additional uncertainty of using estimates for ρ_t and the static parameters.

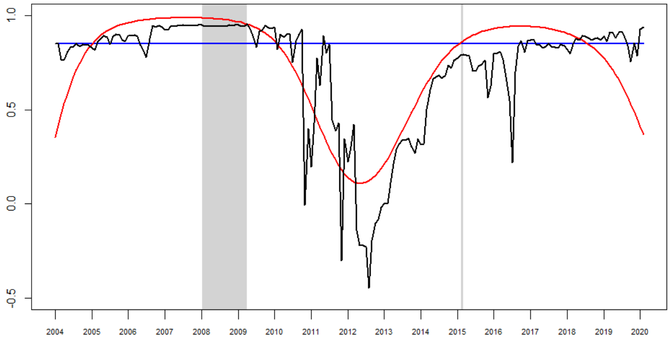
Finally, Figure 3.B.2 compares the estimated correlation, when this is treated as static, to the same estimates obtained with shorter samples. Namely, we use five sub-samples that include observed monthly data form January 2004 up until and including years 2010-2014 (notice that extending the sample with one additional year corresponds to including twelve additional monthly observations). The time-constant correlation model shows a delay of around four years in picking up the deviation between the survey-based data and the claimant counts. This result 1) motivates our choice to model the state correlation as time-varying and 2) further stresses the need for an approach that is able to capture the time-variation of the correlation in a more timely manner, which is achieved by the RBBF rather than the cubic splines method. We therefore conduct a real-time exercise where we estimate the state variables of interest for the sub-samples mentioned above, while estimating the correlation as static and with the RBBF[12]. We compare these results to the ones obtained with the Dutch labour force model that does not include any auxiliary series[13]. Figures 3.B.3-3.B.5 and 3.B.6-3.B.8 respectively show the Kalman filter estimates of the three state variables of interest, and their estimated variances, obtained for all models and sub-samples (as well as for the entire sample, which ends in March 2020)[14]. We notice that, until the end of 2010 and from 2014, the RBBF and the time-constant correlation models both yield close estimates to the ones obtained with the model that does not include auxiliary series. Between these years, however, the latter estimates are much closer to the ones provided by the model which employs the RBBF, rather that the time-constant correlation one. This deviation is due to the incapability of the static correlation estimation method to quickly detect and capture the change in the correlation parameter, contrary to the RBBF. We here use the model without auxiliary series as benchmark, not because we treat is as the true one, but because too large deviations between the state estimates that it yields and the ones obtained with other models, are an indication that some time-varying characteristics are not being taken into account. The variance estimates for $\theta_{y,t}$ and $L_{y,t}$ are always larger, except in the middle of the sample, for the model that does not include auxiliary series as it does not exploit any additional information. The estimation of $R_{y,t}$ is instead affected by a

[12] We do not also perform the real-time analysis with the cubic splines estimation method for the correlation, since the large uncertainty that affects cubic splines at the end of the sample makes this method unsuited for real-time estimation.

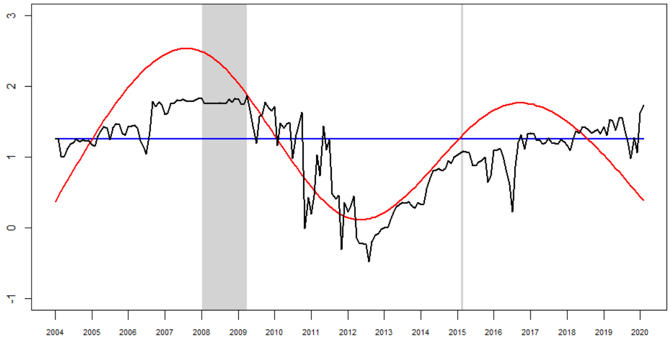
[13] Whenever we add an additional year of observations, we re-estimate the static parameters, by maximum likelihood, of the time-constant correlation model and the one without auxiliary series. The static parameters of the model that employs the RBBF are instead, for computational purposes, always kept equal to the ones obtained by indirect inference for the entire sample (and with knots corresponding to the quartiles of the sample). This choice may have an impact on the results, but we expect it to be marginal given that our simulation study for the ideal RBBF suggests that its performance is not very sensitive to the indirect inference estimates.

[14] Notice that in these Figures the (variance) estimate for the last point in time is not reported when the RBBF is employed, because the algorithm used for the RBBF does not provide such an estimate. This issue is discussed in Section 3.3.2.

much stronger uncertainty, represented by larger variance estimates, when the RBBF is employed for the estimation of the correlation.

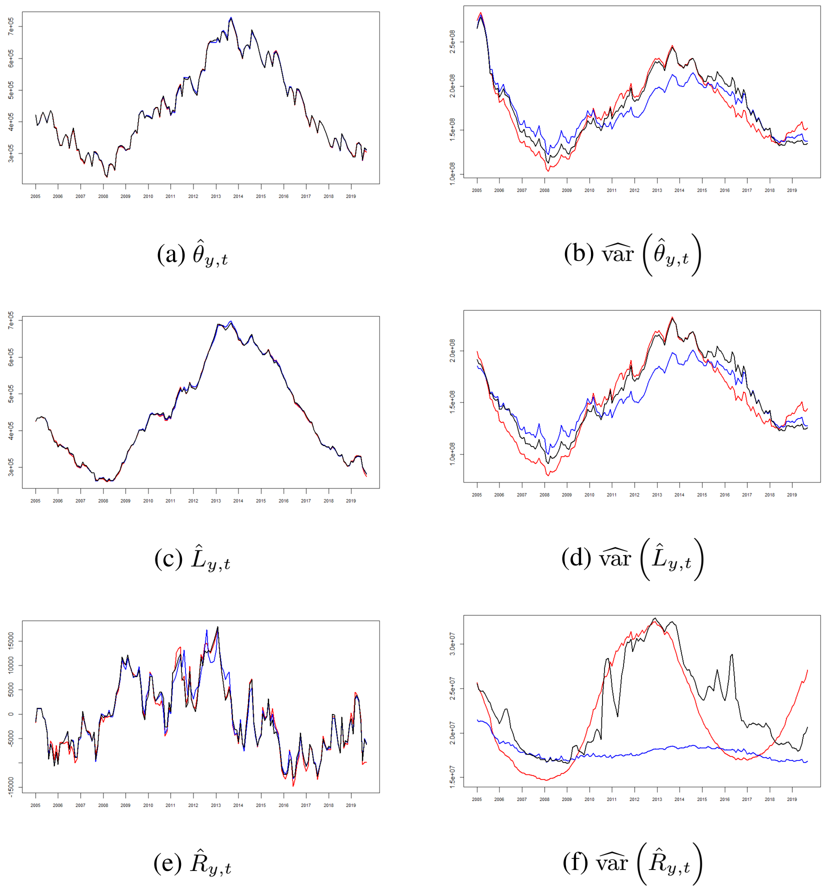


(a) $\hat{\rho}_t$



(b) $\hat{\gamma}_t$

Figure 3.5.1: Time constant (blue), cubic splines (red) and RBBF (black) estimates of ρ_t and γ_t, from the Dutch labour force model extended with the auxiliary series of claimant counts (described in Section 3.2). Monthly data from January 2004 until March 2020 ($T = 195$), $S = 5$, $M = 5000$. The knots for the cubic splines approach correspond to the quartiles of the sample. The first shaded area represent the recession period due to the financial crisis of 2008, whereas the second one refers to the legislative change of February 2015.



(a) $\hat{\theta}_{y,t}$

(b) $\widehat{\text{var}}\left(\hat{\theta}_{y,t}\right)$

(c) $\hat{L}_{y,t}$

(d) $\widehat{\text{var}}\left(\hat{L}_{y,t}\right)$

(e) $\hat{R}_{y,t}$

(f) $\widehat{\text{var}}\left(\hat{R}_{y,t}\right)$

Figure 3.5.2: Kalman filter estimates (left panels) and respective estimated variances (right panels) of the state variables of interest $\theta_{y,t}$, $L_{y,t}$ and $R_{y,t}$, in the Dutch labour force model extended with the auxiliary series of claimant counts (described in Section 3.2). The results are obtained when the correlation parameter is estimated as constant (blue), with the cubic splines method (red), and with the RBBF (black). Monthly data from January 2004 until March 2020 ($T = 195$), $S = 5$, $M = 5000$. The knots for the cubic splines approach correspond to the quartiles of the sample. The first d months are not displayed because of the Kalman filter's diffuse initialisation.

3.6 Conclusions

This chapter proposes a new methodology to estimate nonlinear state space models, where the nonlinearity arises from a stochastic state correlation. The static parameters are estimated with an indirect inference approach, which employs a cubic splines specification for the time-varying correlation as an auxiliary model. The stochastic correlation and remaining state variables are instead estimated with the Rao-Blackwellised bootstrap filter (RBBF). We perform a Monte Carlo simulation study and an empirical application to Dutch unemployment estimation, in order to evaluate the performance of our methodology. In the empirical analysis the correlation represents the relationship between survey-based data about the unemployed labour force, and the series of claimant counts. We compare our method to estimating the time-varying correlation by means of cubic splines only, and as time-constant.

The Monte Carlo simulation study shows that both the cubic splines and the RBBF methods are able to capture the true time-varying pattern of the correlation. This parameter is, on average, more accurately estimated by the former approach. The latter is affected by a strong volatility that tends to deteriorate its performance. This issue also arises from the results of the empirical application. In Section 3.5 we therefore mention two possible solutions that could correct for this problem. Nevertheless, the RBBF beats the cubic splines method in estimating correlations that change rapidly over time, and in yielding more accurate estimates for the correlation towards the end of the sample, especially when sample sizes are small. These latter results already highlight the usefulness of the RBBF in case of real-time estimation.

Both the cubic splines and the RBBF estimators of the correlation are more volatile than the time-constant one, due to their time-varying nature. Hence, in small samples they beat the latter method in terms of squared bias but not in terms of mean squared error (which also captures the volatility of an estimator).

The indirect inference estimators appropriately estimate the static parameters, as their finite-sample distributions are centered around them. Although these distributions are not all symmetrical when sample sizes are small, their shape improves with a larger sample size. So does the performance of the cubic splines and the RBBF estimators, also with respect to estimating the correlation as static.

The point estimates of state variables of interest, when the estimation is not conducted in real-time, do not depend on the method that is employed for the estimation of the time-varying correlation, as much as their estimated variances do. The larger the magnitude of the estimated correlation, the lower the variance estimates of the state variables. In other words, the more information the auxiliary series brings about the variables of interest, the more accurate the estimates for the latter are. Results from the empirical application suggest that the state estimates can instead be rather different when obtained in real time, depending on the method used for the estimation of the correlation. Specifically, the RBBF promptly detects changes in the correlation parameter, thus yielding more reliable state estimates. The static estimation of the correlation parameter, on the other hand, is affected by a strong delay in tackling such changes, which is reflected in unrealistic real-time estimates of the state variables. Moreover, the cubic splines method is not suited for real-time estimation because of its uncertainty in estimating the correlation towards the end of the sample. Real-time estimation of variables is important in the context of official statistics, and finding a method that is reliable for this purpose, such as the RBBF, is an important result.

Empirically, the cubic splines and RBBF agree in estimating a strong and positive correlation in the first and last years of the sample. They capture a deviation between the two types of observed series in the middle of the sample, which is caused by the financial crisis of 2008. The long-term unemployment induced by this recession period can indeed not be completely picked up by the claimant counts. Moreover, only the RBBF manages to tackle an additional drop in the correlation parameter after the implementation of a legislative change that affected the claimant counts series in 2015. This decrease is, however, less protracted and of smaller magnitude than the one triggered by the financial crisis.

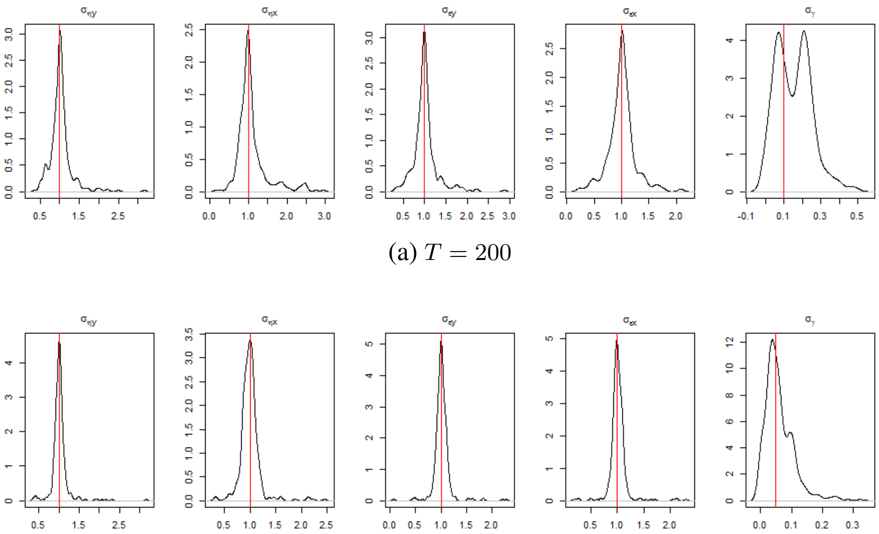
In this chapter we employ only one auxiliary series in the state space model, and hence deal with estimating only one time-varying correlation parameter.

Our proposed method can theoretically be extended to the case where more than one auxiliary series are included in the model, and therefore several correlation parameters need to be estimated. However, the indirect inference and the RBBF are both simulation-based methods and are, as such, computationally rather expensive. Hence, the model should not be too complex in order to guarantee a successful performance of our proposed method.

3.A Additional results from the Monte Carlo simulation study

	$T = 200$						$T = 500$					
	0.9	Sine	Fast sine	Step	Ramp	Random walk	0.9	Sine	Fast sine	Step	Ramp	Random walk
Cubic splines												
MSE($\hat{\rho}_t$)	0.041	0.159	0.206	0.092	0.130	0.124	0.005	0.053	0.116	0.030	0.057	0.051
bias^2($\hat{\rho}_t$)	0.001	0.007	0.081	0.004	0.015	1.4e-04	3.9e-05	0.010	0.080	0.004	0.014	1.1e-04
MSE($\hat{L}_{y,t}$)	0.508	0.597	0.608	0.565	0.592	0.585	0.500	0.584	0.595	0.555	0.583	0.586
bias^2($\hat{L}_{y,t}$)	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Rao-Blackwellised bootstrap filter												
MSE($\hat{\rho}_t$)	0.045	0.172	0.161	0.113	0.172	0.171	0.009	0.129	0.103	0.070	0.129	0.095
bias^2($\hat{\rho}_t$)	0.001	0.050	0.067	0.019	0.037	3.4e-04	1.3e-04	0.030	0.070	0.014	0.028	2.2e-04
MSE($\hat{L}_{y,t}$)	0.946	0.669	0.649	0.746	0.737	0.700	0.527	0.638	0.600	0.602	0.622	0.609
bias^2($\hat{L}_{y,t}$)	0.002	0.002	0.001	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001
Ideal Rao-Blackwellised bootstrap filter												
MSE($\hat{\rho}_t$)						0.159						0.099
bias^2($\hat{\rho}_t$)						3.1e-04						1.3e-04
MSE($\hat{L}_{y,t}$)						0.594						0.596
bias^2($\hat{L}_{y,t}$)						0.001						0.001
Constant												
MSE($\hat{\rho}_t$)	0.002	0.094	0.094	0.072	0.097	0.135	0.001	0.085	0.085	0.067	0.089	0.089
bias^2($\hat{\rho}_t$)	1.7e-05	0.080	0.080	0.063	0.083	2.9e-04	1.4e-06	0.080	0.080	0.062	0.083	0.083
MSE($\hat{L}_{y,t}$)	0.501	0.596	0.599	0.575	0.597	0.596	0.498	0.594	0.593	0.572	0.572	0.594
bias^2($\hat{L}_{y,t}$)	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001

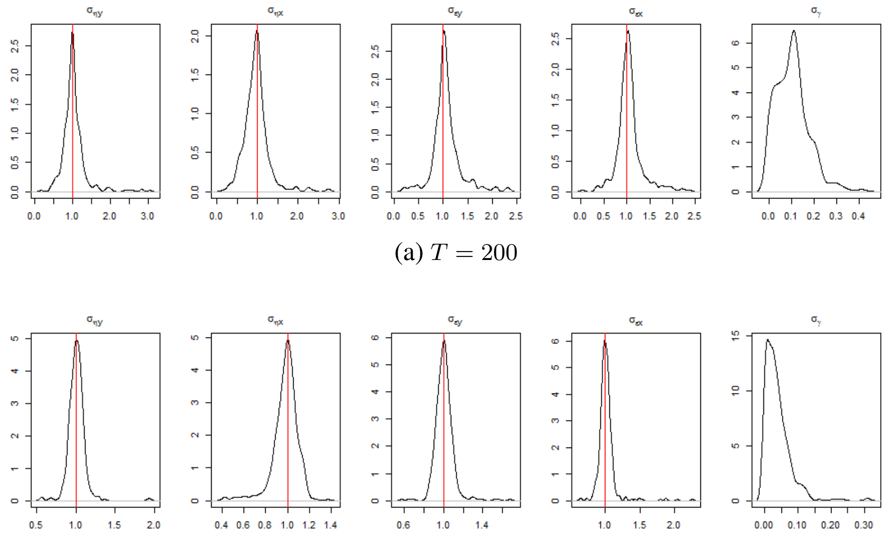
Table 3.A.1: Mean squared error and squared bias for the cubic splines, the Rao-Blackwellised bootstrap filter and the constant estimators of ρ_t, and the Kalman filter estimator of $L_{y,t}$. The second row lists the DGPs for ρ_t. “Ideal Rao-Blackwellised bootstrap filter” indicates that the static parameter vector $\boldsymbol{\tau}$ is treated as known. $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) $T = 200$

(b) $T = 500$

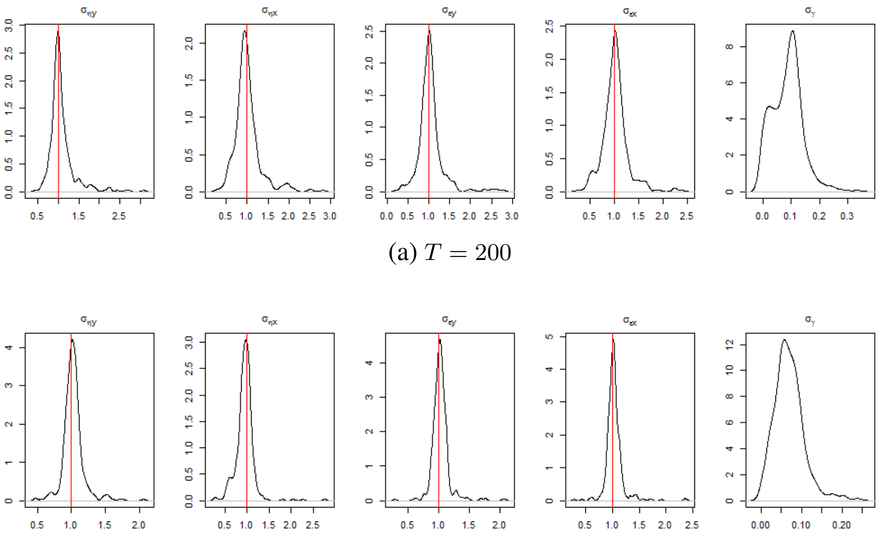
Figure 3.A.1: Distribution of the indirect inference estimators of the static parameters of the nonlinear model, $\boldsymbol{\tau} = (\sigma_{\eta,y}, \sigma_{\eta,x}, \sigma_{\varepsilon,y}, \sigma_{\varepsilon,x}, \sigma_{\gamma})'$, based on the Monte Carlo replicates, when the DGP of ρ_t is a random walk and the BFGS algorithm is not started at the true values; $S = 3$, $n_{\text{sim}} = 500$. The red lines represent the true values of the parameters.



(a) $T = 200$

(b) $T = 500$

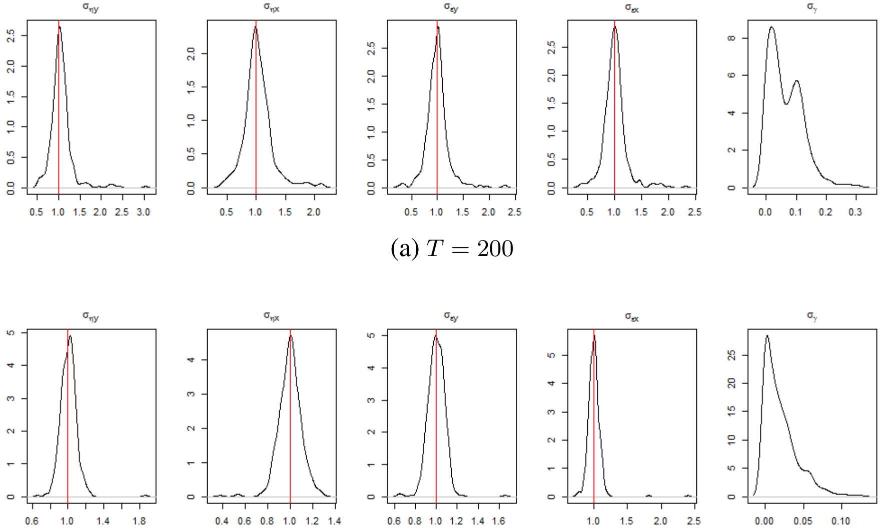
Figure 3.A.2: Distribution of the indirect inference estimators of the static parameters of the nonlinear model, $\boldsymbol{\tau} = (\sigma_{\eta,y}, \sigma_{\eta,x}, \sigma_{\varepsilon,y}, \sigma_{\varepsilon,x}, \sigma_{\gamma})'$, based on the Monte Carlo replicates, when the DGP of ρ_t is constant and equal to 0.9; $S = 3$, $n_{\text{sim}} = 500$. The red lines represent the true values of the parameters.



(a) $T = 200$

(b) $T = 500$

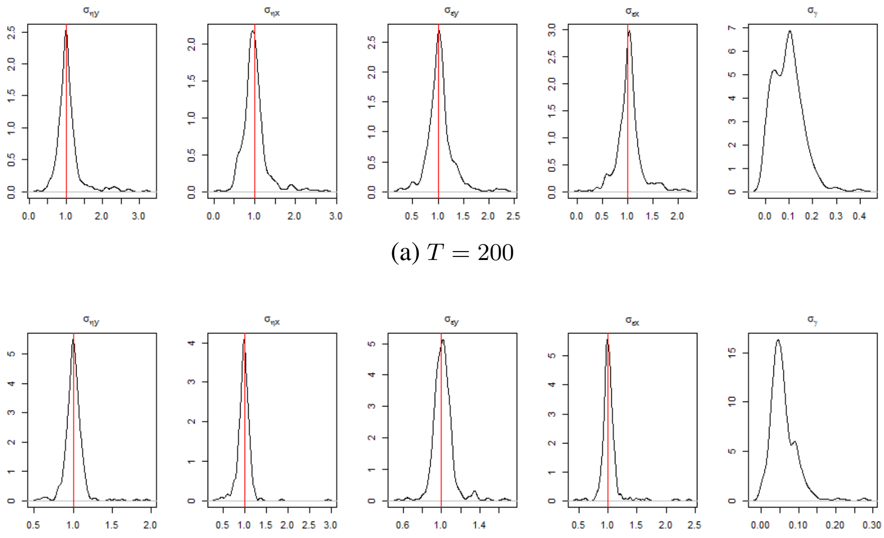
Figure 3.A.3: Distribution of the indirect inference estimators of the static parameters of the nonlinear model, $\boldsymbol{\tau} = (\sigma_{\eta,y}, \sigma_{\eta,x}, \sigma_{\varepsilon,y}, \sigma_{\varepsilon,x}, \sigma_{\gamma})'$, based on the Monte Carlo replicates, when the DGP of ρ_t is a sine function; $S = 3$, $n_{\text{sim}} = 500$. The red lines represent the true values of the parameters.



(a) $T = 200$

(b) $T = 500$

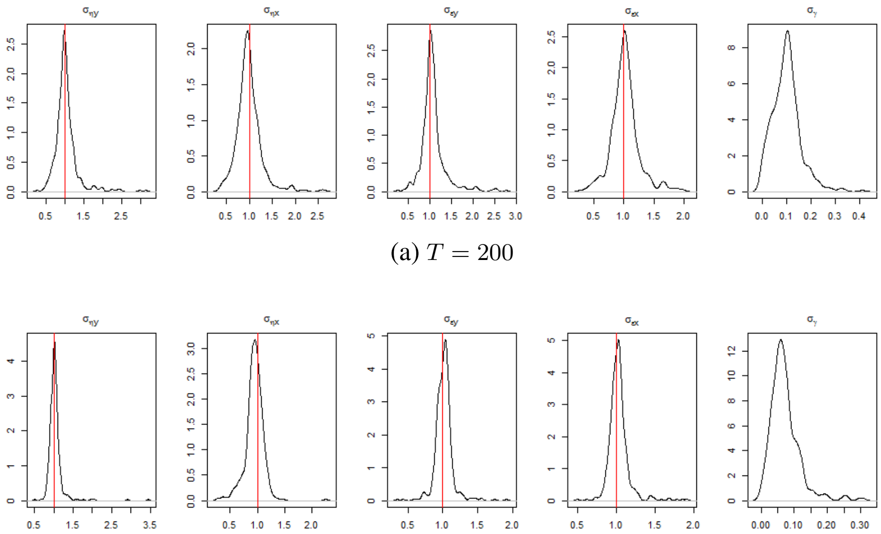
Figure 3.A.4: Distribution of the indirect inference estimators of the static parameters of the nonlinear model, $\boldsymbol{\tau} = (\sigma_{\eta,y}, \sigma_{\eta,x}, \sigma_{\varepsilon,y}, \sigma_{\varepsilon,x}, \sigma_{\gamma})'$, based on the Monte Carlo replicates, when the DGP of ρ_t is a fast sine function; $S = 3$, $n_{\text{sim}} = 500$. The red lines represent the true values of the parameters.



(a) $T = 200$

(b) $T = 500$

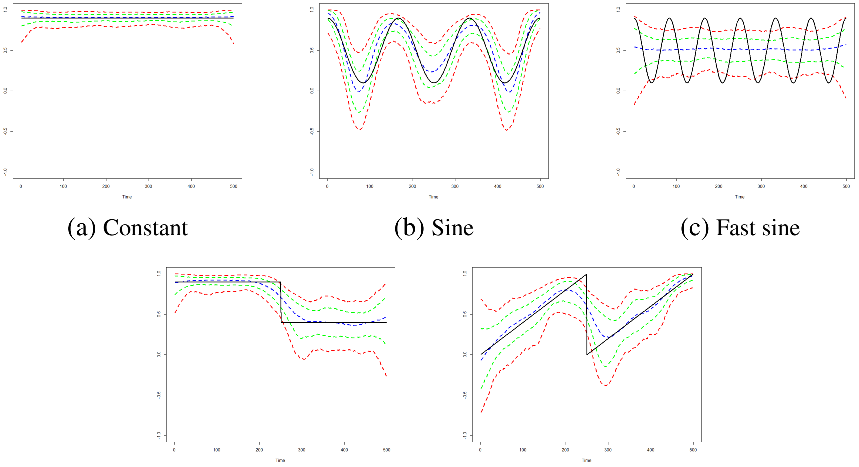
Figure 3.A.5: Distribution of the indirect inference estimators of the static parameters of the nonlinear model, $\boldsymbol{\tau} = (\sigma_{\eta,y}, \sigma_{\eta,x}, \sigma_{\varepsilon,y}, \sigma_{\varepsilon,x}, \sigma_{\gamma})'$, based on the Monte Carlo replicates, when the DGP of ρ_t is a step function; $S = 3$, $n_{\text{sim}} = 500$. The red lines represent the true values of the parameters.



(a) $T = 200$

(b) $T = 500$

Figure 3.A.6: Distribution of the indirect inference estimators of the static parameters of the nonlinear model, $\boldsymbol{\tau} = (\sigma_{\eta,y}, \sigma_{\eta,x}, \sigma_{\varepsilon,y}, \sigma_{\varepsilon,x}, \sigma_{\gamma})'$, based on the Monte Carlo replicates, when the DGP of ρ_t is a ramp function; $S = 3$, $n_{\text{sim}} = 500$. The red lines represent the true values of the parameters.



(a) Constant

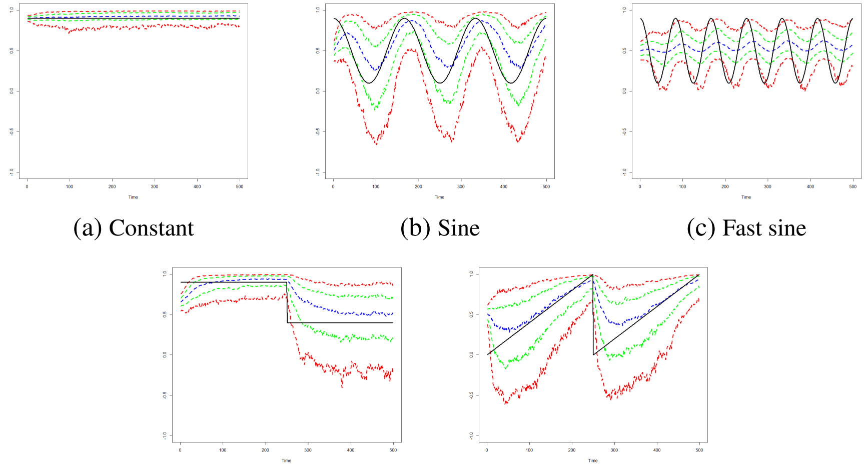
(b) Sine

(c) Fast sine

(d) Step

(e) Ramp

Figure 3.A.7: True process of ρ_t (black) together with the 95% (red), 80% (green) confidence bands, and the median (blue) of the simulation estimates for the cubic splines estimator of ρ_t. $T = 500$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) Constant

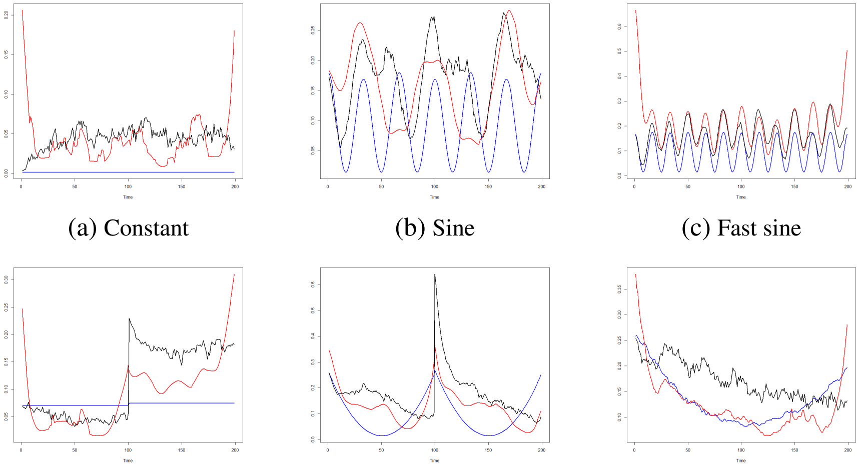
(b) Sine

(c) Fast sine

(d) Step

(e) Ramp

Figure 3.A.8: True process of ρ_t (black) together with the 95% (red), 80% (green) confidence bands, and the median (blue) of the simulation estimates for the Rao-Blackwellised bootstrap filter estimator of ρ_t. $T = 500$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) Constant

(b) Sine

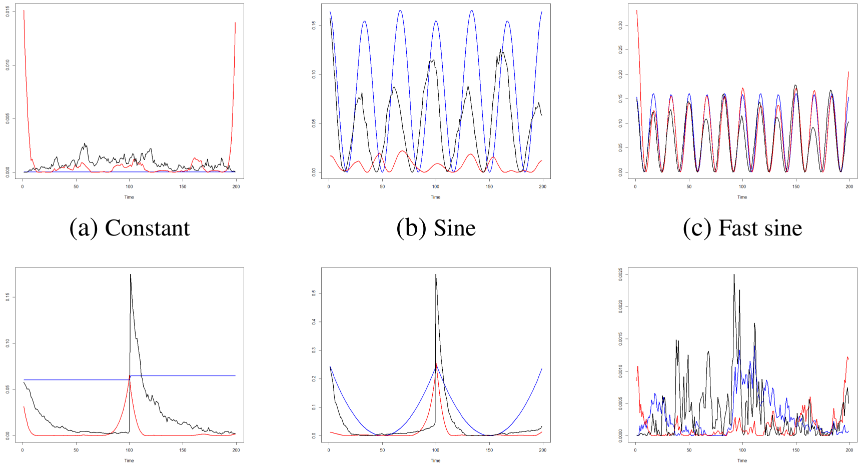
(c) Fast sine

(d) Step

(e) Ramp

(f) Random walk

Figure 3.A.9: MSE of the cubic splines (red), the Rao-Blackwellised bootstrap filter (black) and the constant (blue) estimators of ρ_t, over time. $T = 200$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) Constant

(b) Sine

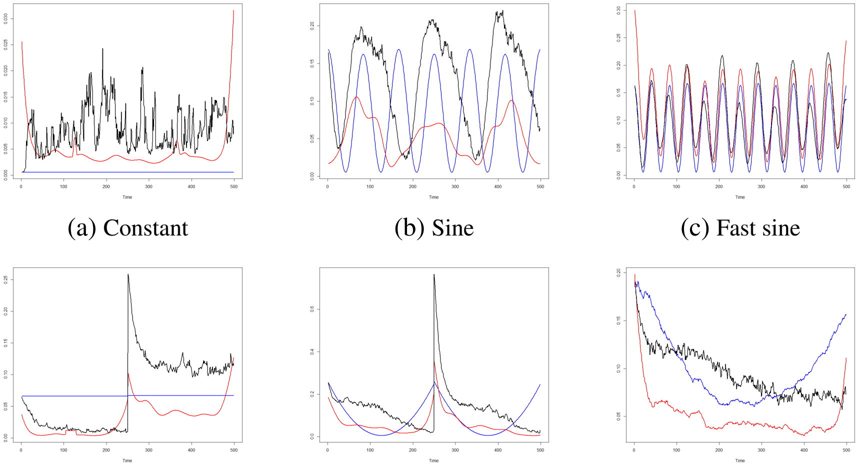
(c) Fast sine

(d) Step

(e) Ramp

(f) Random walk

Figure 3.A.10: Squared bias of the cubic splines (red), the Rao-Blackwellised bootstrap filter (black) and the constant (blue) estimators of ρ_t, over time. $T = 200$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) Constant

(b) Sine

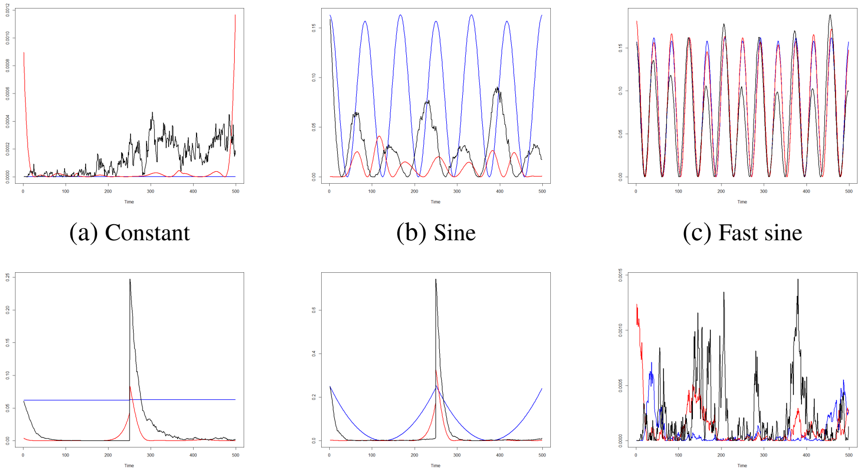
(c) Fast sine

(d) Step

(e) Ramp

(f) Random walk

Figure 3.A.11: MSE of the cubic splines (red), the Rao-Blackwellised bootstrap filter (black) and the constant (blue) estimators of ρ_t, over time. $T = 500$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) Constant

(b) Sine

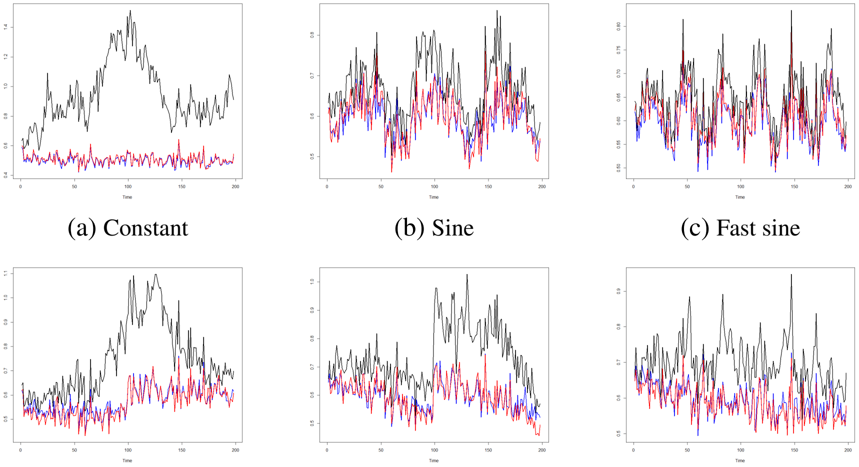
(c) Fast sine

(d) Step

(e) Ramp

(f) Random walk

Figure 3.A.12: Squared bias of the cubic splines (red), the Rao-Blackwellised bootstrap filter (black) and the constant (blue) estimators of ρ_t, over time. $T = 500$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) Constant

(b) Sine

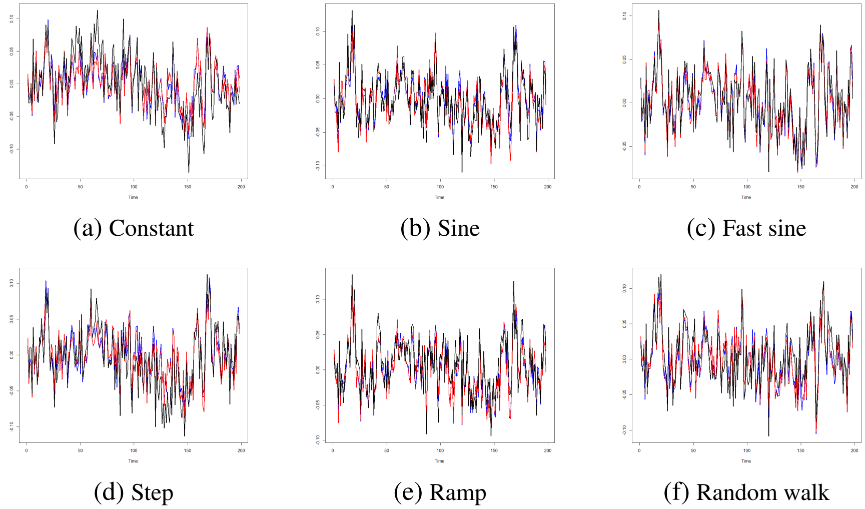
(c) Fast sine

(d) Step

(e) Ramp

(f) Random walk

Figure 3.A.13: MSE, over time, of the Kalman filter estimator of $L_{y,t}$ when ρ_t is estimated by the cubic splines (red) and the Rao-Blackwellised bootstrap filter (black) methods, and as time-constant (blue). $T = 200$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) Constant

(b) Sine

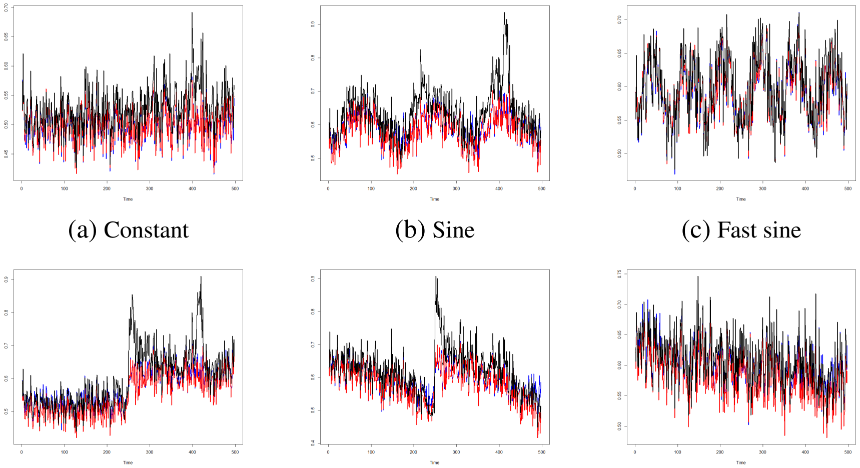
(c) Fast sine

(d) Step

(e) Ramp

(f) Random walk

Figure 3.A.14: Squared bias, over time, of the Kalman filter estimator of $L_{y,t}$ when ρ_t is estimated by the cubic splines (red) and the Rao-Blackwellised bootstrap filter (black) methods, and as time-constant (blue). $T = 200$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) Constant

(b) Sine

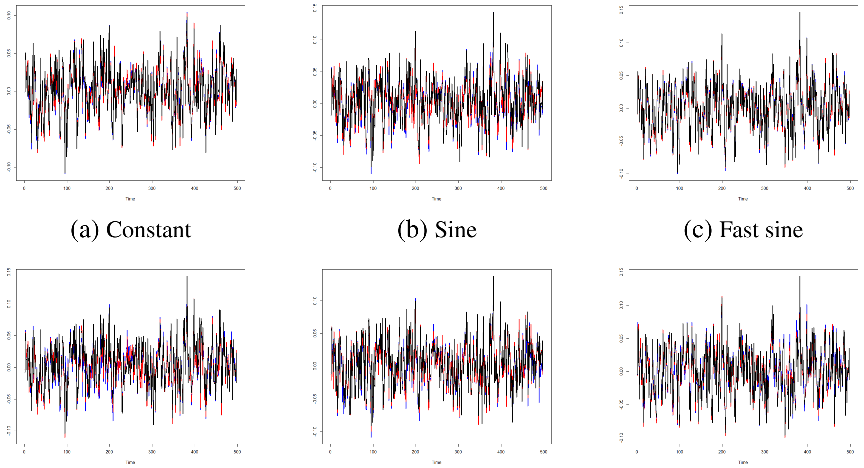
(c) Fast sine

(d) Step

(e) Ramp

(f) Random walk

Figure 3.A.15: MSE, over time, of the Kalman filter estimator of $L_{y,t}$ when ρ_t is estimated by the cubic splines (red) and the Rao-Blackwellised bootstrap filter (black) methods, and as time-constant (blue). $T = 500$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.



(a) Constant

(b) Sine

(c) Fast sine

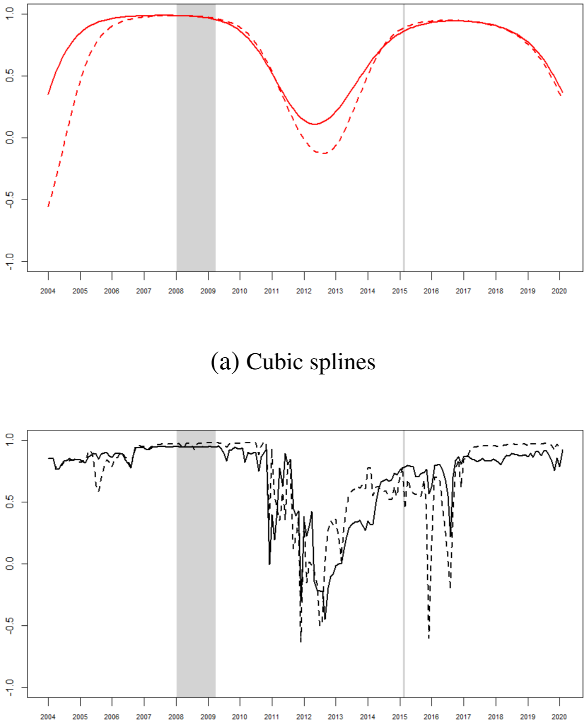
(d) Step

(e) Ramp

(f) Random walk

Figure 3.A.16: Squared bias, over time, of the Kalman filter estimator of $L_{y,t}$ when ρ_t is estimated by the cubic splines (red) and the Rao-Blackwellised bootstrap filter (black) methods, and as time-constant (blue). $T = 500$, $S = 3$, $M = 5000$, $n_{\text{sim}} = 500$.

3.B Additional results from the empirical application



(a) Cubic splines

(b) RBBF

Figure 3.B.1: Cubic splines and RBBF estimates of ρ_t, from the Dutch labour force model extended with the auxiliary series of claimant counts (described in Section 3.2). Monthly data from January 2004 until March 2020 ($T = 195$), $S = 5$, $M = 5000$. The dashed lines refer to the setting where the knots for the cubic splines approach are approximately equally distant by fixing February 2015 as one of them, otherwise they correspond to the quartiles of the sample (the solid lines are the same as in Figure 3.5.1a). The first shaded area represent the recession period due to the financial crisis of 2018, whereas the second one refers to the legislative change of February 2015.

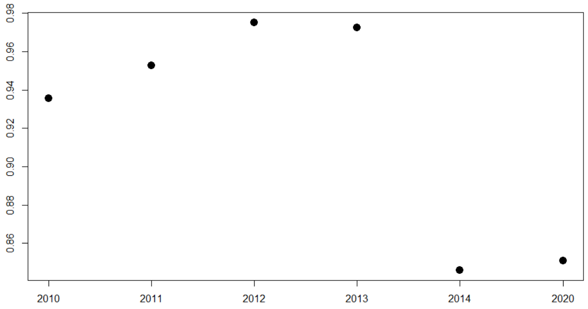
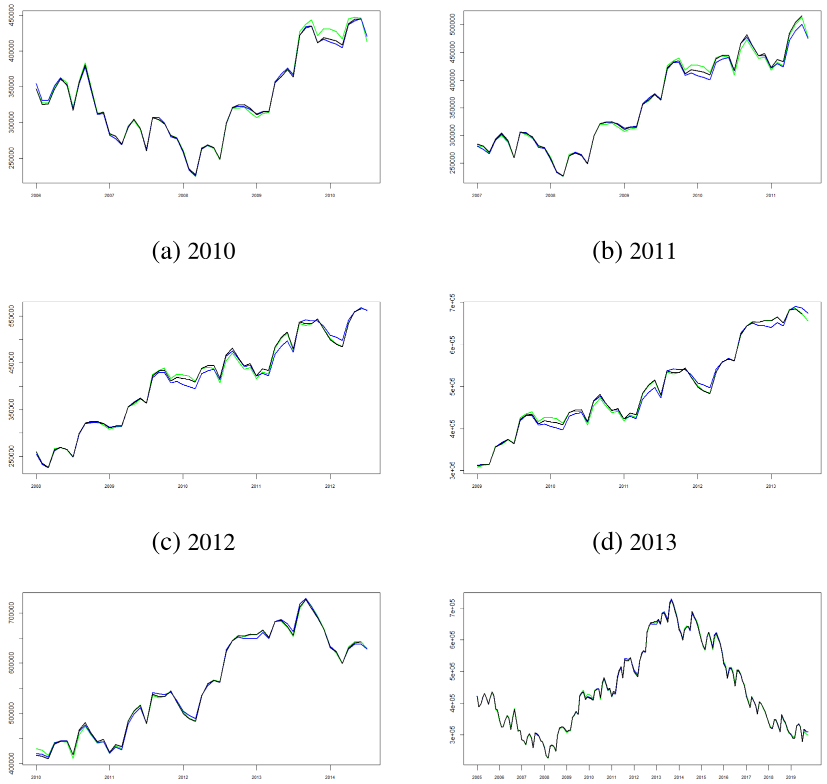


Figure 3.B.2: Estimates of the correlation parameter when this is treated as time-constant, obtained with monthly data observed from January 2004 up to and including the year displayed on the horizontal axis. The results refer to the Dutch labour force model extended with the auxiliary series of claimant counts (described in Section 3.2).



(a) 2010

(b) 2011

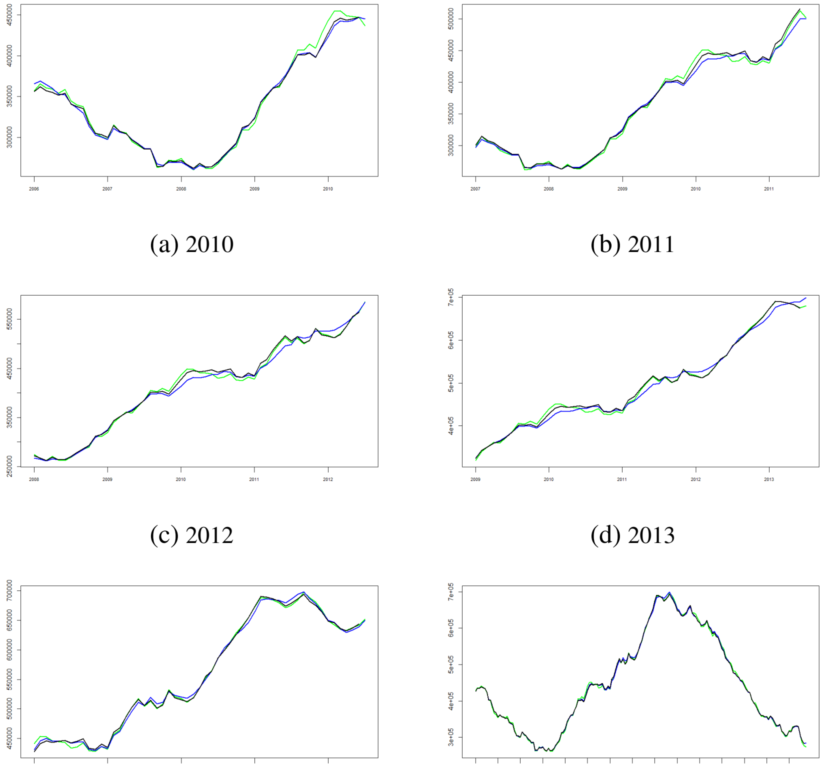
(c) 2012

(d) 2013

(e) 2014

(f) 2020

Figure 3.B.3: Kalman filter estimates of $\theta_{y,t}$ in the Dutch labour force model (described in Section 3.2). The green lines refer to the model without auxiliary series. The blue and black lines refer to the model extended with the auxiliary series of claimant counts, when the correlation is estimated as time constant and with the RBBF, respectively. Each panel shows the results obtained with monthly data observed from January 2004 up to and including the year displayed in the respective caption. We do not always show estimates for all time periods in order to facilitate the comparison among panels.



(a) 2010

(b) 2011

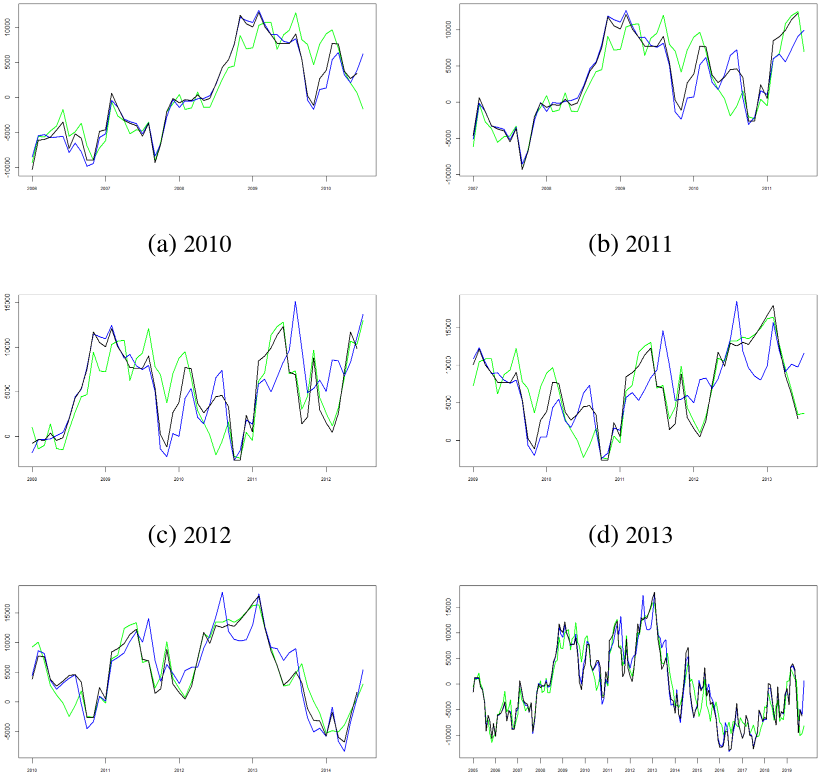
(c) 2012

(d) 2013

(e) 2014

(f) 2020

Figure 3.B.4: Kalman filter estimates of $L_{y,t}$ in the Dutch labour force model (described in Section 3.2). The green lines refer to the model without auxiliary series. The blue and black lines refer to the model extended with the auxiliary series of claimant counts, when the correlation is estimated as time constant and with the RBBF, respectively. Each panel shows the results obtained with monthly data observed from January 2004 up to and including the year displayed in the respective caption. We do not always show estimates for all time periods in order to facilitate the comparison among panels.



(a) 2010

(b) 2011

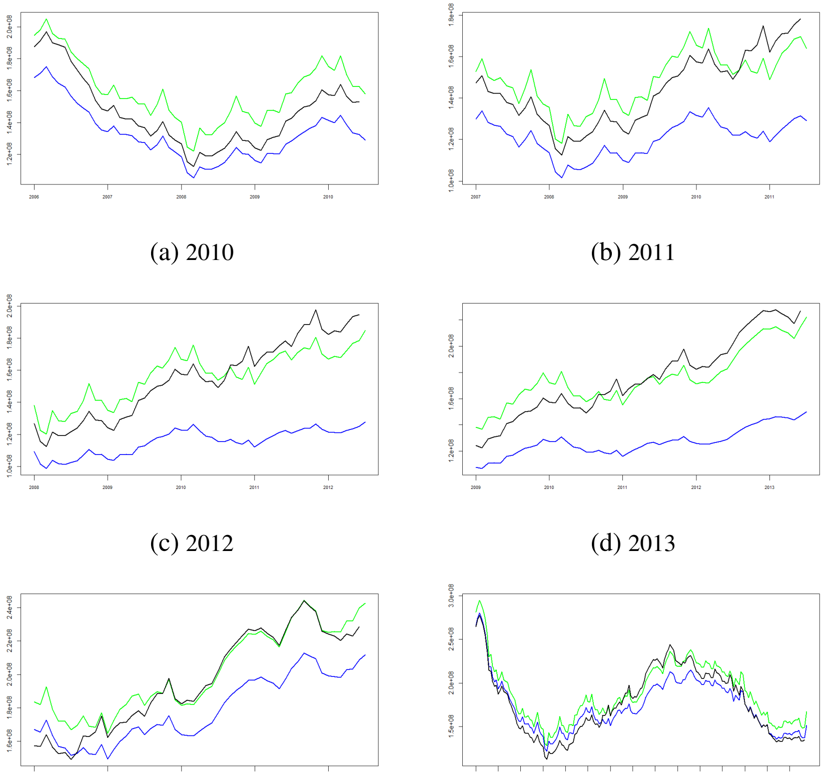
(c) 2012

(d) 2013

(e) 2014

(f) 2020

Figure 3.B.5: Kalman filter estimates of $R_{y,t}$ in the Dutch labour force model (described in Section 3.2). The green lines refer to the model without auxiliary series. The blue and black lines refer to the model extended with the auxiliary series of claimant counts, when the correlation is estimated as time constant and with the RBBF, respectively. Each panel shows the results obtained with monthly data observed from January 2004 up to and including the year displayed in the respective caption. We do not always show estimates for all time periods in order to facilitate the comparison among panels.



(a) 2010

(b) 2011

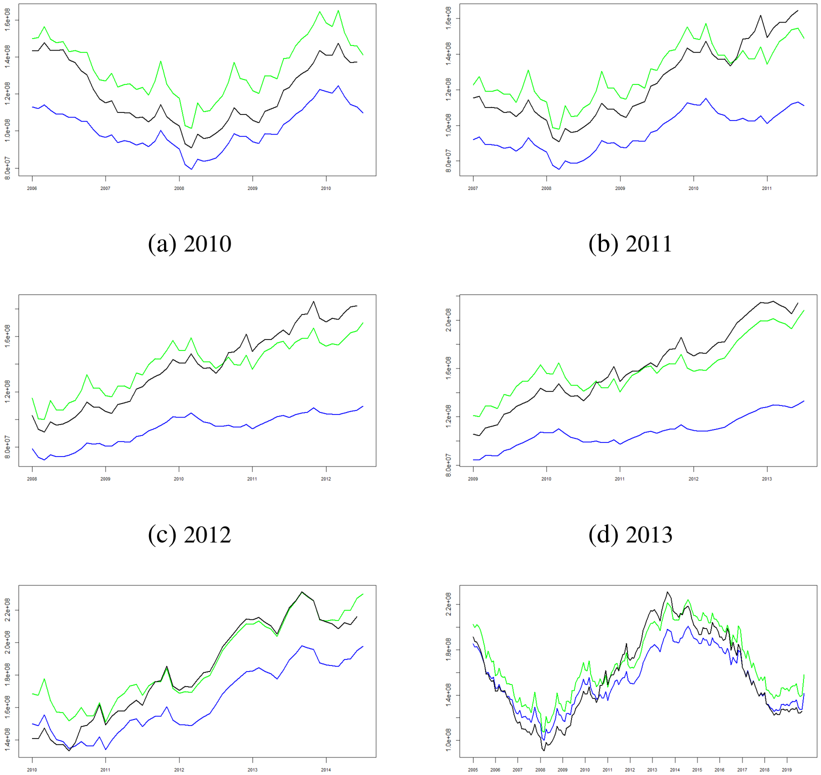
(c) 2012

(d) 2013

(e) 2014

(f) 2020

Figure 3.B.6: Estimated variances, $\widehat{\text{var}}\left(\hat{\theta}_{y,t}\right)$, of the Kalman filter estimates of $\theta_{y,t}$ in the Dutch labour force model (described in Section 3.2). The green lines refer to the model without auxiliary series. The blue and black lines refer to the model extended with the auxiliary series of claimant counts, when the correlation is estimated as time constant and with the RBBF, respectively. Each panel shows the results obtained with monthly data observed from January 2004 up to and including the year displayed in the respective caption. We do not always show estimates for all time periods in order to facilitate the comparison among panels.



(a) 2010

(b) 2011

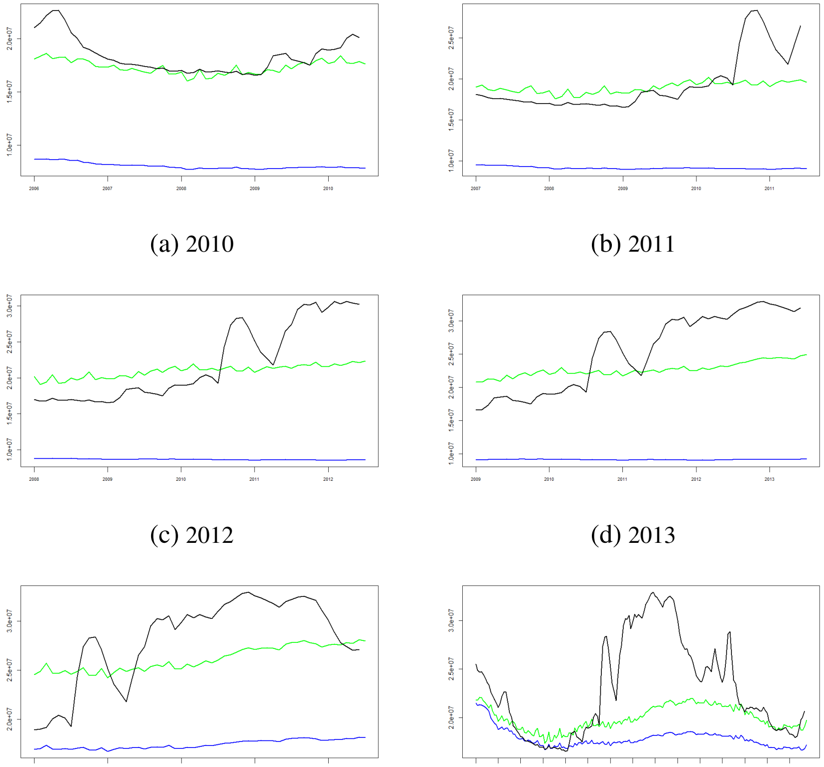
(c) 2012

(d) 2013

(e) 2014

(f) 2020

Figure 3.B.7: Estimated variances, $\widehat{\text{var}}\left(\hat{L}_{y,t}\right)$, of the Kalman filter estimates of $L_{y,t}$ in the Dutch labour force model (described in Section 3.2). The green lines refer to the model without auxiliary series. The blue and black lines refer to the model extended with the auxiliary series of claimant counts, when the correlation is estimated as time constant and with the RBBF, respectively. Each panel shows the results obtained with monthly data observed from January 2004 up to and including the year displayed in the respective caption. We do not always show estimates for all time periods in order to facilitate the comparison among panels.



(a) 2010

(b) 2011

(c) 2012

(d) 2013

(e) 2014

(f) 2020

Figure 3.B.8: Estimated variances, $\widehat{\text{var}}\left(\hat{R}_{y,t}\right)$, of the Kalman filter estimates of $R_{y,t}$ in the Dutch labour force model (described in Section 3.2). The green lines refer to the model without auxiliary series. The blue and black lines refer to the model extended with the auxiliary series of claimant counts, when the correlation is estimated as time constant and with the RBBF, respectively. Each panel shows the results obtained with monthly data observed from January 2004 up to and including the year displayed in the respective caption. We do not always show estimates for all time periods in order to facilitate the comparison among panels.

4

On score-driven, extended Kalman filter and importance sampling methods, and their attempt to estimate time-varying state correlations

Abstract

Part of doing research involves taking directions that do not lead anywhere or that yield unsuccessful results. Hence, what is discovered during such journeys is often left unpublished. A PhD thesis is, on the other hand, an opportunity to share also negative results, as well as documenting in more detail all the work that is behind a project. In this chapter we therefore discuss other econometric methods, than the ones presented in Chapter 3, that we, unsuccessfully, try to use in order to estimate time-varying state correlations in state space models. Specifically, we consider a score-driven approach, the extended Kalman filter and importance sampling. The former method does not yield a satisfactory performance, compared to the cubic splines approach adopted in Chapter 3. We also cannot find a way to feasibly implement the latter two methods. All these issues are discussed in this chapter.

4.1 Introduction

This chapter is made of a collection of notes that have been taken by the author while exploring other econometric methods, than the ones presented in Chapter 3, in order to tackle the same problem: estimating time-varying state correlations in state space models. Such methods did not yield a successful or feasible performance, as will be shown, and have therefore not been included in Chapter 3. However, we believe that presenting negative results can improve science and we therefore dedicate this chapter to them.

We first consider another method, than the cubic splines discussed in Chapter 3, that deals with a deterministic time-variation of the correlation parameters, and that is here presented in order to deal with multiple state correlations in the state space model (not only one as we did, for simplicity, in Chapter 3). This method is called the generalized autoregressive score (GAS) method and has been proposed by Delle Monache et al. (2016) in order to estimate time-varying parameters in state space models. The approach assumes that the time-varying correlations depend on their past values as well as past values of the (scaled) score of the log-likelihood (which is required in order to estimate the state space model). Closed form expressions for the scores are needed and derived in this chapter. The scores are supposed to provide the direction, at each point in time, for updating the time-varying correlations in order to improve the model's local fit, but the GAS estimates of the correlations look more volatile than the estimates obtained with the cubic splines approach.

As in Chapter 3, we also investigate two additional estimation methods that instead assume that the correlations are varying stochastically over time. The latter parameters are therefore treated as additional unobserved components and the state space model in this case becomes nonlinear. The nonlinearity of the state space model hampers its estimation with the standard technique of Kalman filtering. The idea underlying both approaches is to apply the Kalman filter to a linear model that approximates the nonlinear one as much as possible (and better than the cubic splines model). The first method is the extended Kalman filter and consists of linearising the model based on a first-order Taylor expansion around the estimates of the unobserved components (Durbin & Koopman, 2012, Chapter 10). It is possible to then estimate the approximate linear state space model with the standard Kalman filter. The second approach is based on importance sampling and involves a linear model that approximates the nonlinear one up to the second order. As such, the approximation is more accurate than the one employed by the extended Kalman filter, but the estimation method becomes more complex. The Kalman filter can again be applied to the approximate linear model, and the estimation accuracy of the unobserved components can be further improved by Monte Carlo simulation methods (Durbin & Koopman, 2012, Chapter 11). For simplicity, we here show how these two methodologies can potentially deal with the estimation of only one state correlation parameter.

This chapter is purely methodological since we explain how the three abovementioned methods can attempt to estimate rather simple state space models, not the (extended) Dutch labour force model of Chapters 2 and 3. For the GAS method only, we report also the results of a Monte Carlo simulation study which demonstrates that this approach is not as successful as the cubic splines one. We will finally show that the linearisation required by the extended Kalman filter and importance sampling methods is non-trivial to obtain, thus rendering these approaches (or at least the way we tried to implement them) impractical to estimate time-varying state correlations in state space models.

We re-introduce the general form of the state space model (equation (3.3.1) of Section 3.3), for a $n \times 1$ observable vector $\boldsymbol{z}_t = (\boldsymbol{y}_t', \boldsymbol{x}_t')'$:

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{Z}\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N\left(\boldsymbol{0}, \boldsymbol{H}\right) \\ \boldsymbol{\alpha}_t &= \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{R}\boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N\left(\boldsymbol{0}, \boldsymbol{Q}_t\right), \end{aligned} \tag{4.1.1}$$

for $t = 1, \ldots, T$, where T is the sample size. The $m \times 1$ vector $\boldsymbol{\alpha}_t$ contains the state variables and $\boldsymbol{\eta}_t$ is the corresponding $r \times 1$ vector of disturbances. The respective $r \times r$ covariance matrix, $\boldsymbol{Q}_t$, is assumed to have time-varying correlation parameter(s). The $m \times m$ matrix $\boldsymbol{T}$ defines the dynamic structure of the state vector, whereas the $n \times m$ matrix $\boldsymbol{Z}$ links the observed series to the latent variables. The $n \times 1$ vector $\boldsymbol{\varepsilon}_t$ represents the innovation term of the observation equation. The selection matrix $\boldsymbol{R}$ has dimensions $m \times r$.

4.2 Score-driven method

In the score-driven framework, we let $\boldsymbol{Q}_t$ contain k correlation parameters, which we collect in the $k \times 1$ vector $\boldsymbol{\rho}_t$. We further assume the matrices $\boldsymbol{Z}$, $\boldsymbol{H}$, $\boldsymbol{T}$ and $\boldsymbol{Q}_t$ to be non-stochastic. The score-driven approach indeed employs a specification for the time-varying correlations that preserves the linearity of the state space model, as the time-varying correlations do not depend on components that are random at time t, given past information. Let $\boldsymbol{Z}_{t-1} = \{\boldsymbol{z}_{t-1}, \ldots, \boldsymbol{z}_1\}$ and $\boldsymbol{\mathcal{P}}_t = \{\boldsymbol{\rho}_t, \ldots, \boldsymbol{\rho}_1\}$. We define the available information set at time t as $\boldsymbol{\mathcal{F}}_t = \{\boldsymbol{Z}_{t-1}, \boldsymbol{\mathcal{P}}_t\}$. Conditional on the information set and on the vector of time-constant parameters, $\boldsymbol{\beta}$ (i.e, the time-constant parameters contained in $\boldsymbol{H}$, $\boldsymbol{T}$ and $\boldsymbol{Q}_t$, which are also referred to as "hyperparameters"), the observations and the state vector are Gaussian: $\boldsymbol{z}_t|\boldsymbol{\mathcal{F}}_t;\boldsymbol{\beta} \sim N\left(\boldsymbol{Z}\boldsymbol{a}_t, \boldsymbol{F}_t\right)$ and $\boldsymbol{\alpha}_t|\boldsymbol{\mathcal{F}}_t;\boldsymbol{\beta} \sim N\left(\boldsymbol{a}_t, \boldsymbol{P}_t\right)$. Therefore, the log-likelihood function for $\boldsymbol{z}_t$ at time t takes the form:

$$\ell_t = \log p\left(\boldsymbol{z}_t|\boldsymbol{\mathcal{F}}_t;\boldsymbol{\beta}\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{1}{2}\log\left(\det \boldsymbol{F}_t\right) - \frac{1}{2}\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t, \quad (4.2.1)$$

where the prediction error $\boldsymbol{v}_t$ and its covariance matrix $\boldsymbol{F}_t$ are estimated by means of the following standard Kalman filter recursions:

$$\begin{aligned} \boldsymbol{v}_t &= \boldsymbol{z}_t - \boldsymbol{Z}\boldsymbol{a}_t \\ \boldsymbol{F}_t &= \boldsymbol{Z}\boldsymbol{P}_t\boldsymbol{Z}' + \boldsymbol{H} \\ \boldsymbol{K}_t &= \boldsymbol{T}\boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1} \\ \boldsymbol{a}_{t|t} &= \boldsymbol{a}_t + \boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t \end{aligned} \qquad \begin{aligned} \boldsymbol{P}_{t|t} &= \boldsymbol{P}_t - \boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\boldsymbol{Z}\boldsymbol{P}_t \\ \boldsymbol{a}_{t+1} &= \boldsymbol{T}\boldsymbol{a}_t + \boldsymbol{K}_t\boldsymbol{v}_t \\ \boldsymbol{P}_{t+1} &= \boldsymbol{T}\boldsymbol{P}_t\left(\boldsymbol{T} - \boldsymbol{K}_t\boldsymbol{Z}\right)' + \boldsymbol{R}\boldsymbol{Q}_t\boldsymbol{R}', \end{aligned} \quad (4.2.2)$$

for $t = 1, \ldots, T$. The vector $\boldsymbol{a}_{t|t}$ represents the filtered estimate of $\boldsymbol{\alpha}_t$, and $\boldsymbol{P}_{t|t}$ is its filtered estimated covariance matrix. The one-step ahead forecast (based on $\boldsymbol{\mathcal{F}}_t$) of the state vector is $\boldsymbol{a}_t$, and the corresponding forecasted covariance matrix is $\boldsymbol{P}_t$. The recursions for the Kalman smoother (KFS) use also future information to estimate the state variables, and can be found in Durbin and Koopman (2012, Chapter 4).

For observation-driven methods, and specifically for the GAS method, originally proposed by Creal et al. (2013) and Harvey (2013), the time-varying parameters are assumed to depend on deterministic functions of their lagged values as well as lagged values of the (scaled) scores of the log-likelihood. The latter are the first derivatives of the log-likelihood with respect to the time-varying parameters. Since such lagged values are assumed to be known at the current time, the time-varying parameters are perfectly predictable. Moreover, the log-likelihood is available in closed form. The scores provide the directions for updating the time-varying parameters in order to improve the model's local fit in terms of the log-likelihood. Since these updates occur at each point in time, the estimates of the time-varying correlations are expected to be non-smooth.

We specify the following GAS, with orders p and q, updating equation for the time-varying vector $\boldsymbol{\rho}_t$:

$$\boldsymbol{\rho}_{t+1} = \boldsymbol{\tau} + \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{s}_{t-j+1} + \sum_{j=1}^{q} \boldsymbol{B}_j \boldsymbol{\rho}_{t-i+1}, \quad t = 1, \ldots, T, \tag{4.2.3}$$

where $\boldsymbol{\tau}$ is a $k \times 1$ vector of constants and $\boldsymbol{A}_i$ and $\boldsymbol{B}_j$ are diagonal coefficient matrices. The scaled-score of the conditional log-likelihood is $\boldsymbol{s}_t = \boldsymbol{\mathcal{S}}_t \nabla_t$, where

$$\nabla_t = \frac{\partial \ell_t}{\partial \boldsymbol{\rho}_t}, \quad \boldsymbol{\mathcal{S}}_t \equiv \boldsymbol{\mathcal{I}}_t^{-l} = \left[\mathrm{E}_{t-1} \left(\nabla_t \nabla_t' \right) \right]^{-l}, \quad l = 0, 1/2, 1.$$

Under a correct model specification, $\boldsymbol{\mathcal{S}}_t = \left[- \mathrm{E}_{t-1} \left(\frac{\partial^2 \ell_t}{\partial \boldsymbol{\rho}_t \partial \boldsymbol{\rho}_t'} \right) \right]^{-l}$ and it can be proven that the score function $\boldsymbol{s}_t$ forms a martingale difference sequence: $\mathrm{E}_{t-1} \left[\boldsymbol{s}_t \right] = \mathbf{0}$. This follows from the information matrix equality and the properties of the score vector (zero mean score/zero expected gradient). Moreover, if we choose $\boldsymbol{\mathcal{S}}_t = \boldsymbol{\mathcal{I}}_t^{-1/2}$, then $\boldsymbol{s}_t$ has unit variance, since $\mathrm{E}_{t-1} \left[\boldsymbol{s}_t \boldsymbol{s}_t' \right] = \boldsymbol{S}_t \boldsymbol{\mathcal{I}}_t \boldsymbol{S}_t'$. If we express $\boldsymbol{\rho}_t$ in terms of its infinite moving average representation, then we need the roots of the equation $|\boldsymbol{I} - \boldsymbol{B}(L)| = \mathbf{0}$, where L is the lag operator, to lie inside the unit circle in order for $\boldsymbol{\rho}_t$ to be covariance stationary.

For a GAS(1,1) process, the infinite moving-average representation is

$$\boldsymbol{\rho}_t = (\boldsymbol{I} - \boldsymbol{B}_1)^{-1}\boldsymbol{\tau} + \boldsymbol{A}_1 \sum_{i=0}^{\infty} \boldsymbol{B}_1^i \boldsymbol{s}_{t-i}. \tag{4.2.4}$$

In the Monte Carlo simulation study we always use $l = \frac{1}{2}$, in line with previous literature. We avoid numerical instability by replacing $\boldsymbol{\mathcal{I}}_t$ with the smoothed estimator $\tilde{\boldsymbol{\mathcal{I}}}_t = (1-\zeta)\boldsymbol{\mathcal{I}}_t + \zeta\tilde{\boldsymbol{\mathcal{I}}}_{t-1}$, where ζ is estimated by maximising the log-likelihood, with $\zeta \in [0,1]$. Notice that the vector of time-constant parameters, $\boldsymbol{\beta}$, contains ζ, the time-invariant hyperparameters of model (4.1.1), as well as $\boldsymbol{\tau}$ and the coefficient matrices of equation (4.2.3).

Delle Monache et al. (2016) show that in the context of a state space model with time-varying parameters, the gradient and the information matrix, respectively, take the form

$$\begin{aligned} \nabla_t &= \frac{1}{2}\left[\dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\operatorname{vec}\left(\boldsymbol{v}_t\boldsymbol{v}_t' - \boldsymbol{F}_t\right) - 2\dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right] \\ \boldsymbol{\mathcal{I}}_t &= \frac{1}{2}\left[\dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{F}}_t + 2\dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\dot{\boldsymbol{V}}_t\right], \end{aligned} \tag{4.2.5}$$

for $t = 1, \dots, T$, where $\dot{\boldsymbol{V}}_t = \frac{\partial \boldsymbol{v}_t}{\partial \boldsymbol{\rho}_t'}$ and $\dot{\boldsymbol{F}}_t = \frac{\partial \operatorname{vec}(\boldsymbol{F}_t)}{\partial \boldsymbol{\rho}_t'}$ are computed via the following additional recursions:

$$\begin{aligned}
\dot{\boldsymbol{V}}_t &= -\boldsymbol{Z}\dot{\boldsymbol{A}}_t \\
\dot{\boldsymbol{F}}_t &= (\boldsymbol{Z}\otimes\boldsymbol{Z})\,\dot{\boldsymbol{P}}_t \\
\dot{\boldsymbol{K}}_t &= \left(\boldsymbol{F}_t^{-1}\boldsymbol{Z}\otimes\boldsymbol{T}\right)\dot{\boldsymbol{P}}_t - \left(\boldsymbol{F}_t^{-1}\otimes\boldsymbol{K}_t\right)\dot{\boldsymbol{F}}_t \\
\dot{\boldsymbol{A}}_{t|t} &= \dot{\boldsymbol{A}}_t + \left(\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{Z}\otimes\boldsymbol{I}_m\right)\dot{\boldsymbol{P}}_t - \left(\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\otimes\boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{F}}_t \\
&\quad + \boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\dot{\boldsymbol{V}}_t \\
\dot{\boldsymbol{P}}_{t|t} &= \dot{\boldsymbol{P}}_t - \left[\boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\boldsymbol{Z}\otimes\boldsymbol{I}_m + \boldsymbol{I}_m\otimes\boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\boldsymbol{Z}\right]\dot{\boldsymbol{P}}_t \\
&\quad + \left(\boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\otimes\boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{F}}_t \\
\dot{\boldsymbol{A}}_{t+1} &= \boldsymbol{T}\dot{\boldsymbol{A}}_t + \left(\boldsymbol{v}_t'\otimes\boldsymbol{I}_m\right)\dot{\boldsymbol{K}}_t + \boldsymbol{K}_t\dot{\boldsymbol{V}}_t \\
\dot{\boldsymbol{P}}_{t+1} &= \left[(\boldsymbol{T}\otimes\boldsymbol{T}) - (\boldsymbol{K}_t\boldsymbol{Z}\otimes\boldsymbol{T})\right]\dot{\boldsymbol{P}}_t - \left(\boldsymbol{I}_m\otimes\boldsymbol{T}\boldsymbol{P}_t\boldsymbol{Z}'\right)\boldsymbol{C}_{mn}\dot{\boldsymbol{K}}_t \\
&\quad + (\boldsymbol{R}\otimes\boldsymbol{R})\,\dot{\boldsymbol{Q}}_t,
\end{aligned} \tag{4.2.6}$$

for $t = 1,\ldots,T$, with $\dot{\boldsymbol{Q}}_t = \frac{\partial\,\mathrm{vec}(\boldsymbol{Q}_t)}{\partial\boldsymbol{\rho}_t'}$. The $mn \times mn$ commutation matrix $\boldsymbol{C}_{mn}$ is such that $\boldsymbol{C}_{mn}\,\mathrm{vec}\,(\boldsymbol{X}) = \mathrm{vec}\,(\boldsymbol{X}')$, given a $m \times n$ matrix $\boldsymbol{X}$. The $\mathrm{vec}\,(\boldsymbol{X})$ operator stacks the columns of $\boldsymbol{X}$, one underneath the other. The above-listed additional recursions are based on the assumption that $\boldsymbol{Q}_t$ is the only matrix containing time-varying parameters.

The derivations of equations (4.2.5) and (4.2.6) follow, with minor differences, the ones of Delle Monache et al. (2016), and are reported in Appendix 4.A.

We need to impose a restriction on the space of $\boldsymbol{\rho}_t$, namely of being between -1 and 1. We therefore re-parametrize the correlation parameter vector as

$$\boldsymbol{\rho}_t = \tanh\left(\boldsymbol{\gamma}_t\right),$$

where $\tanh$ is the hyperbolic tangent, which is a time-invariant, continuous, invertible, and twice differentiable function. The vector $\boldsymbol{\gamma}_t$, instead of $\boldsymbol{\rho}_t$, now follows the updating rule of equation (4.2.3). The equations that define the gradient and the Fisher information matrix in formula (4.2.5) therefore have to take the Jacobian $\frac{\partial \boldsymbol{\rho}_t}{\partial \boldsymbol{\gamma}_t'}$ into account, by taking the form

$$\begin{aligned} \nabla_t &= \frac{\partial \ell_t}{\partial \boldsymbol{\gamma}_t} = \left[\frac{\partial \ell_t}{\partial \boldsymbol{\gamma}_t'} \right]' = \left[\frac{\partial \ell_t}{\partial \boldsymbol{\rho}_t'} \frac{\partial \boldsymbol{\rho}_t}{\partial \boldsymbol{\gamma}_t'} \right]' = \frac{\partial \boldsymbol{\rho}_t'}{\partial \boldsymbol{\gamma}_t} \frac{\partial \ell_t}{\partial \boldsymbol{\rho}_t}, \\ \boldsymbol{\mathcal{I}}_t &= -\operatorname{E}_{t-1} \left(\frac{\partial^2 \ell_t}{\partial \boldsymbol{\gamma}_t \partial \boldsymbol{\gamma}_t'} \right) = -\operatorname{E}_{t-1} \left(\frac{\partial \boldsymbol{\rho}_t'}{\partial \boldsymbol{\gamma}_t} \frac{\partial^2 \ell_t}{\partial \boldsymbol{\rho}_t \partial \boldsymbol{\rho}_t'} \frac{\partial \boldsymbol{\rho}_t}{\partial \boldsymbol{\gamma}_t'} \right) \\ &= -\frac{\partial \boldsymbol{\rho}_t'}{\partial \boldsymbol{\gamma}_t} \operatorname{E}_{t-1} \left(\frac{\partial^2 \ell_t}{\partial \boldsymbol{\rho}_t \partial \boldsymbol{\rho}_t'} \right) \frac{\partial \boldsymbol{\rho}_t}{\partial \boldsymbol{\gamma}_t'}. \end{aligned} \tag{4.2.7}$$

In our case the matrix $\boldsymbol{Q}_t$, containing the time-varying correlations, takes the form

$$\boldsymbol{Q}_t = \begin{bmatrix} \sigma_1^2 & \rho_{t,1,2}\sigma_1\sigma_2 & \cdots & \rho_{t,1,m}\sigma_1\sigma_m \\ \rho_{t,1,2}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{t,2,m}\sigma_2\sigma_m \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{t,1,m}\sigma_1\sigma_m & \rho_{t,2,m}\sigma_2\sigma_m & \cdots & \sigma_m^2 \end{bmatrix}.$$

Therefore, $\dot{\boldsymbol{Q}}_t = \frac{\partial \operatorname{vec}(\boldsymbol{Q}_t)}{\partial \boldsymbol{\rho}_t'}$ is a $m^2 \times k$ matrix and $\frac{\partial \boldsymbol{\rho}_t}{\partial \boldsymbol{\gamma}_t'} = \operatorname{diag}\left(1 - \tanh^2(\gamma_{t,1,2}), \ldots, 1 - \tanh^2(\gamma_{t,m-1,m})\right)$ is a $k \times k$ diagonal matrix[1]. Consider, as illustration, the following simple bivariate example:

$$\boldsymbol{Q}_t = \begin{bmatrix} \sigma_1^2 & \rho_t\sigma_1\sigma_2 \\ \rho_t\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

$$\text{Then, } \frac{\partial \operatorname{vec}(\boldsymbol{Q}_t)}{\partial \rho_t} = \begin{pmatrix} 0 \\ \sigma_1\sigma_2 \\ \sigma_1\sigma_2 \\ 0 \end{pmatrix} \text{ and } \frac{\partial \rho_t}{\partial \gamma_t} = 1 - \tanh^2(\gamma_t). \tag{4.2.8}$$

[1] If the parametrization $\boldsymbol{\rho}_t = \frac{\gamma_t}{\sqrt{1+\gamma_t^2}}$ is used, then $\frac{\partial \boldsymbol{\rho}_t}{\partial \boldsymbol{\gamma}_t'} = \operatorname{diag}\left(1/\left(1+\gamma_{t,1,2}^2\right)^{\frac{3}{2}}, \ldots, 1/\left(1+\gamma_{t,m-1,m}^2\right)^{\frac{3}{2}}\right)$. Different parametrizations may be employed to simplify mathematical derivations, and to avoid numerical problems.

The orders of the updating equation (4.2.3), p and q, can be chosen by information criteria. The Akaike and Bayesian information criteria for state space models with diffuse initialization take the form (Durbin & Koopman, 2012, Chapter 7):

$$\begin{aligned} \text{AIC} &= \frac{1}{T}\left[-2\ell_d + 2\left(d + \dim(\boldsymbol{\beta})\right)\right] \\ \text{BIC} &= \frac{1}{T}\left[-2\ell_d + \log\left(T\right)\left(d + \dim(\boldsymbol{\beta})\right)\right], \end{aligned} \tag{4.2.9}$$

where d is the number of nonstationary state variables of $\boldsymbol{y}_t$, $\dim(\boldsymbol{\beta})$ is the dimension of $\boldsymbol{\beta}$, and ℓ_d is the diffuse log-likelihood (Harvey, 1989):

$$\ell_d = -\frac{Tn}{2}\log(2\pi) - \frac{1}{2}\sum_{t=d+1}^{T}\left[\log\left(\det \boldsymbol{F}_t\right) - \frac{1}{2}\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right]. \tag{4.2.10}$$

However, we constrain our analysis to $p = q = 1$ since the GAS(1,1) model is already very flexible due to its infinite moving average representation (4.2.4), but restricted by the “ARMA” structure.

All the time-constant parameters, $\boldsymbol{\beta}$, are estimated by maximum likelihood.

4.2.1 Monte Carlo simulation study

The Monte Carlo simulation study is conducted in order to evaluate the performance of the GAS method in estimating a time-varying state correlation. Such performance is compared to the one of the cubic splines approach discussed in Chapter 3, and to estimating the correlation parameter as time constant, i.e., by maximum likelihood. We consider the following state space model specifications, which entail only one state correlation parameter to be estimated:

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{Z}\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N\left(\boldsymbol{0}, \boldsymbol{H}\right) \\ \boldsymbol{\alpha}_t &= \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N\left(\boldsymbol{0}, \boldsymbol{Q}_t\right), \quad t = 1, \ldots, T. \end{aligned}$$

1. Bivariate local level model: $\boldsymbol{z}_t$, $\boldsymbol{\alpha}_t$, $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are 2×1 vectors. $\boldsymbol{\alpha}_t = (L_{1,t}, L_{2,t})'$, $\boldsymbol{Z} = \boldsymbol{T} = \boldsymbol{I}_2$, $\boldsymbol{H} = \text{diag}(1, 1)$ and

$$\boldsymbol{Q}_t \quad = \quad \begin{bmatrix} \exp(2\sigma_1) & \tanh(\gamma_t)\exp(\sigma_1)\exp(\sigma_2) \\ \tanh(\gamma_t)\exp(\sigma_1)\exp(\sigma_2) & \exp(2\sigma_2) \end{bmatrix},$$

with $\sigma_1 = \sigma_2 = 0$.

2. Bivariate smooth trend model: $\boldsymbol{z}_t$, $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are 2×1 vectors; $\boldsymbol{\alpha}_t = (L_{1,t}, R_{1,t}, L_{2,t}, R_{2,t})'$ is a 4×1 vector. $\boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$,

$$\boldsymbol{T} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{H} = \mathrm{diag}(1,1) \text{ and}$$

$$\boldsymbol{Q}_t = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \exp(2\sigma_1) & 0 & \tanh(\gamma_t)\exp(\sigma_1)\exp(\sigma_2) \\ 0 & 0 & 0 & 0 \\ 0 & \tanh(\gamma_t)\exp(\sigma_1)\exp(\sigma_2) & 0 & \exp(2\sigma_2) \end{bmatrix},$$

with $\sigma_1 = \sigma_2 = 0$ (which implies $\exp(\sigma_1) = \exp(\sigma_2) = 1$).

For both model specifications and all estimation methods, we consider the following data generating processes (DGPs) for the time-varying parameter ρ_t, which are similar to the ones employed in Chapter 3:

1. No-correlation: $\rho_t = 0$
2. Constant correlation: $\rho_t = 0.9$
3. Smoothly decreasing function: $\rho_t = 0.9/(1 + \exp(0.02(t - T/2))$
4. Sine: $\rho_t = 0.5 + 0.4\cos(2\pi t/(T/5))$
5. Fast sine: $\rho_t = 0.5 + 0.4\cos(2\pi t/(T/50))$
6. Step: $\rho_t = 0.9 - 0.5(t > T/2)$
7. Ramp: $\rho_t = 2/T \mod (t/(T/2))$

for $t = 1, \ldots, T$. We consider two different sample sizes, $T = 200$ and $T = 500$, and run $n_{\text{sim}} = 500$ Monte Carlo simulations.

Figures 4.2.1 and 4.2.2 display the 5%, 10%, 80%, and 95% percentiles (which we loosely call confidence bands), and the median, of the simulation estimates of the correlation parameter obtained with the GAS and the cubic splines methods, respectively. The results refer to the bivariate local level model when $T = 200$. The GAS approach barely manages to follow the true time-varying patterns of ρ_t, and its confidence bands are rather wide, which is an indication of a weak accuracy of the estimation method. The cubic splines technique instead shows a better performance in terms of estimation accuracy and capability of capturing the true time-variation (or constancy) of the correlation parameter. Figures 4.A.3 and 4.A.4 report the same results for the larger sample size of $T = 500$. Although, as expected, the estimation accuracy of both methods improves, the cubic splines approach confirms its superior performance. Similar conclusions can be drawn for the bivariate smooth trend model, as Figures 4.A.7-4.A.10 show. These results indicate that the cubic splines are a better choice, than the GAS, for a deterministic specification of the time-varying correlation, and subsequent approximate linear model, employed in the indirect inference estimation of Chapter 3. Finally, figures 4.A.1,4.A.2, 4.A.5 and 4.A.6 display the same results discussed above, when the correlation parameter is estimated as time constant.

In Section 3.1 we wrote: “Solving our problem, i.e., extracting this time-varying state correlation, is therefore already challenging by the fact that this parameter relates innovations of components that are unobserved. Additionally, data and their respective log-likelihood functions, are much less informative about correlations than other parameters, such as means or variances.”. This is of course an issue for all estimation methods. However, our intuition behind the failure of the score-driven method is that the additional Kalman filter recursions needed for it, make the relationship between the score-driven estimator of the correlation parameter and the observed data highly complex, and so render further difficult to extract the relevant (for us) information from the data. Both the cubic spline (maximum likelihood) and the bootstrap filter estimators, instead, depend on the data “only” via the Kalman filter recursions (not the additional ones).

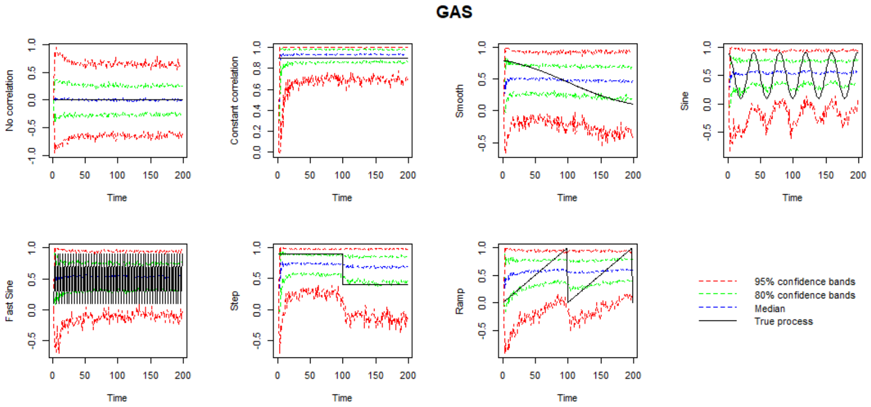


Figure 4.2.1: Bivariate local level model. $T = 200$, $n_{\text{sim}} = 500$.

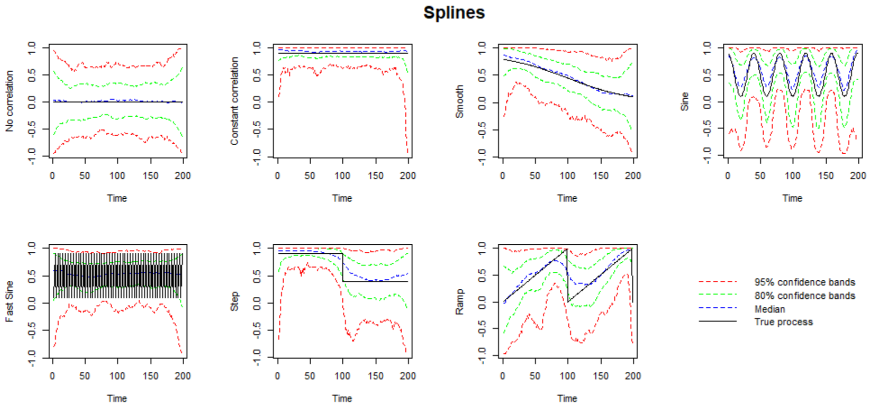


Figure 4.2.2: Bivariate local level model. $T = 200$, $n_{\text{sim}} = 500$.

4.3 Extended Kalman filter

In this Section we explain how to implement the extended Kalman filter for a bivariate local level model with time-varying level correlations (Appendix 4.B shows how to implement the extended Kalman filter for a bivariate smooth trend model).

As already mention in the Introduction, we now make the assumption of a stochastic state correlation, which is therefore treated as an additional unobserved component in the model.

The observation equations for a bivariate local level model are:

$$\begin{aligned} y_t &= \alpha_t^y + \varepsilon_t^y \\ x_t &= \alpha_t^x + \varepsilon_t^x, \quad t = 1, \ldots, T, \end{aligned}$$

with $(\varepsilon_t^y, \varepsilon_t^x)' \sim N(\mathbf{0}, \boldsymbol{H})$.

The transition equations are:

$$\begin{aligned} \alpha_{t+1}^y &= \alpha_t^y + \eta_t^y \\ \alpha_{t+1}^x &= \alpha_t^x + \eta_t^x \\ \gamma_{t+1} &= \gamma_t + \eta_t^\gamma, \quad t = 1, \ldots, T, \end{aligned}$$

with $(\eta_t^y, \eta_t^x, \eta_t^\gamma)' \sim N(\mathbf{0}, \boldsymbol{Q}_t)$, and $\boldsymbol{Q}_t = \begin{bmatrix} \sigma_y^2 & \tanh(\gamma_t)\sigma_y\sigma_x & 0 \\ \tanh(\gamma_t)\sigma_y\sigma_x & \sigma_x^2 & 0 \\ 0 & 0 & \sigma_\gamma^2 \end{bmatrix}$.

The bivariate local level model can be written in compact notation as equation (4.1.1), with $\boldsymbol{z}_t = (x_t, y_t)'$, $\boldsymbol{\varepsilon}_t = (\varepsilon_t^y, \varepsilon_t^x)'$, $\boldsymbol{\alpha}_t = (\alpha_t^y, \alpha_t^x, \gamma_t)'$, $\boldsymbol{\eta}_t = (\eta_t^y, \eta_t^x)'$, $\boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, $\boldsymbol{T} = \boldsymbol{I}_3$, and $\boldsymbol{R} = \boldsymbol{I}_2$.

In this case γ_t is an additional state variable with a random walk transition equation and its own source of error, η_t^γ. The state space model therefore becomes nonlinear since the correlation parameter in the covariance matrix $\boldsymbol{Q}_t$, is random at time t. The extended Kalman filter can be used in order to linearise the model, if the nonlinearity is present among state variables. In this case the non-linearity is caused by the interaction of the state variable γ_t with the innovation vector $\boldsymbol{\eta}_t$. Therefore, we first transform the model by means of the Cholesky decomposition of $\boldsymbol{Q}_t$: $\boldsymbol{Q}_t = \boldsymbol{C}_t\boldsymbol{C}_t'$, with $\boldsymbol{C}_t = \begin{bmatrix} \sigma_y & 0 & 0 \\ \tanh(\gamma_t)\sigma_x & \sigma_x\sqrt{1-\tanh^2(\gamma_t)} & 0 \\ 0 & 0 & \sigma_\gamma \end{bmatrix}$. We then replace $\boldsymbol{\alpha}_t$ by $\boldsymbol{C}_t\boldsymbol{\alpha}_t^*$ in the observation equation of model (4.1.1):

$$\boldsymbol{z}_t = \boldsymbol{Z}\boldsymbol{C}_t\boldsymbol{\alpha}_t^* + \boldsymbol{\varepsilon}_t = \boldsymbol{Z}\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim (\boldsymbol{0}, \boldsymbol{H}), \quad t = 1, \ldots, T,$$

where $\boldsymbol{\alpha}_t = \boldsymbol{C}_t\boldsymbol{\alpha}_t^* = \begin{bmatrix} \sigma_y & 0 & 0 \\ \tanh(\gamma_t)\sigma_x & \sigma_x\sqrt{1-\tanh^2(\gamma_t)} & 0 \\ 0 & 0 & \sigma_\gamma \end{bmatrix} \begin{pmatrix} \alpha_t^{y*} \\ \alpha_t^{x*} \\ \gamma_t \end{pmatrix}$.

The observation equations of the transformed local level model are:

$$\begin{aligned} y_t &= \sigma_y\alpha_t^{y*} + \varepsilon_t^y \\ x_t &= \sigma_x\left(\tanh(\gamma_t)\alpha_t^{y*} + \sqrt{1-\tanh^2(\gamma_t)}\alpha_t^{x*}\right) + \varepsilon_t^x, \quad t = 1, \ldots, T, \end{aligned} \tag{4.3.1}$$

with $(\varepsilon_t^y, \varepsilon_t^x)' \sim N(\boldsymbol{0}, \boldsymbol{H})$.

The transition equations of the transformed model are:

$$\begin{aligned} \alpha_{t+1}^{y*} &= \alpha_t^{y*} + \eta_t^{y*} \\ \alpha_{t+1}^{x*} &= \alpha_t^{x*} + \eta_t^{x*} \\ \gamma_{t+1} &= \gamma_t + \eta_t^\gamma, \quad t = 1, \ldots, T, \end{aligned}$$

with $\left(\eta_t^{y*}, \eta_t^{x*}, \eta_t^\gamma\right)' \sim N(\boldsymbol{0}, \boldsymbol{Q}^*)$ and $\boldsymbol{Q}^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma_\gamma^2 \end{bmatrix}$.

Notice that the elements of the transformed innovation vector $\left(\eta_t^{y*}, \eta_t^{x*}, \eta_t^{\gamma}\right)'$ are not correlated anymore, and that therefore their covariance matrix, $\boldsymbol{Q}^*$, is now time-constant.

The nonlinearity is now present in the observation equation (4.3.1), which can be written as:

$$\boldsymbol{z}_t = Z_t(\boldsymbol{\alpha}_t^*) + \boldsymbol{\varepsilon}_t, \tag{4.3.2}$$

where

$$Z_t(\boldsymbol{\alpha}_t^*) = \begin{pmatrix} \sigma_y \alpha_t^{y*} \\ \sigma_x \left(\tanh(\gamma_t)\alpha_t^{y*} + \sqrt{1-\tanh^2(\gamma_t)}\alpha_t^{x*}\right) \end{pmatrix}.$$

The extended Kalman filter linearises the model by approximating $Z_t(\boldsymbol{\alpha}_t^*)$ with its first-order Taylor expansion around $\boldsymbol{a}_t$ (i.e., the predicted state vector):

$$Z_t(\boldsymbol{\alpha}_t^*) \approx Z_t(\boldsymbol{a}_t) + \dot{\boldsymbol{Z}}_t(\boldsymbol{\alpha}_t^* - \boldsymbol{a}_t),$$

where

$$\begin{aligned} \dot{\boldsymbol{Z}}_t &= \left.\frac{\partial Z_t(\boldsymbol{\alpha}_t^*)}{\partial \boldsymbol{\alpha}_t^{*'}}\right|_{\boldsymbol{\alpha}_t^*=\boldsymbol{a}_t} \\ &= \left.\frac{\partial Z_t(\boldsymbol{\alpha}_t^*)}{\partial\left(\alpha_t^{y*}, \alpha_t^{x*}, \gamma_t\right)}\right|_{\boldsymbol{\alpha}_t^*=\boldsymbol{a}_t} \\ &= \left[\begin{matrix} \sigma_y & 0 \\ \sigma_x \tanh(\gamma_t) & \sigma_x\sqrt{1-\tanh^2(\gamma_t)} \end{matrix}\right. \\ &\qquad \left.\left.\begin{matrix} 0 \\ \sigma_x\sqrt{1-\tanh^2(\gamma_t)}\left(\sqrt{1-\tanh^2(\gamma_t)}\alpha_t^{y*} - \tanh(\gamma_t)\alpha_t^{x*}\right) \end{matrix}\right]\right|_{\boldsymbol{\alpha}_t^*=\boldsymbol{a}_t}. \end{aligned}$$

The extended Kalman filter recursions, needed to estimate the state variables, are:

$$\begin{aligned} \boldsymbol{v}_t &= \boldsymbol{z}_t - Z_t(\boldsymbol{a}_t) \\ \boldsymbol{F}_t &= \dot{\boldsymbol{Z}}_t \boldsymbol{P}_t \dot{\boldsymbol{Z}}_t{}' + \boldsymbol{H} \\ \boldsymbol{K}_t &= \boldsymbol{T} \boldsymbol{P}_t \dot{\boldsymbol{Z}}_t{}' \boldsymbol{F}_t^{-1} \\ \boldsymbol{a}_{t|t} &= \boldsymbol{a}_t + \boldsymbol{P}_t \dot{\boldsymbol{Z}}_t{}' \boldsymbol{F}_t^{-1} \boldsymbol{v}_t \end{aligned} \qquad \begin{aligned} \boldsymbol{P}_{t|t} &= \boldsymbol{P}_t - \boldsymbol{P}_t \dot{\boldsymbol{Z}}_t{}' \boldsymbol{F}_t^{-1} \dot{\boldsymbol{Z}}_t \boldsymbol{P}_t \\ \boldsymbol{a}_{t+1} &= \boldsymbol{T} \boldsymbol{a}_t + \boldsymbol{K}_t \boldsymbol{v}_t \\ \boldsymbol{P}_{t+1} &= \boldsymbol{T} \boldsymbol{P}_t \left(\boldsymbol{T} - \boldsymbol{K}_t \dot{\boldsymbol{Z}}_t \right)' + \boldsymbol{R} \boldsymbol{Q}^* \boldsymbol{R}', \end{aligned}$$

for $t = 1, \ldots, T$. The recursions for the extended state smoother are:

$$\begin{aligned} \boldsymbol{r}_{t-1} &= \dot{\boldsymbol{Z}}_t' \boldsymbol{F}_t^{-1} \boldsymbol{v}_t + \left(\boldsymbol{T} - \boldsymbol{K}_t \dot{\boldsymbol{Z}}_t \right)' \boldsymbol{r}_t \\ \hat{\boldsymbol{\alpha}}_t &= \boldsymbol{a}_t + \boldsymbol{P}_t \boldsymbol{r}_{t-1}, \end{aligned}$$

for $t = T, \ldots, 1$, and with $\boldsymbol{r}_T = \mathbf{0}$.

The filtered and smoothed estimates for γ_t correspond, respectively, to the last element of $\boldsymbol{a}_{t|t}$ and $\hat{\boldsymbol{\alpha}}_t$, and the confidence intervals can be built using the corresponding diagonal element of $\boldsymbol{P}_{t|t}$, for $t = 1, \ldots, T$. By taking the hyperbolic tangent of the estimate and the confidence intervals, we find the respective estimate and confidence intervals for ρ_t.

4.4 Importance sampling

In this Section we show how to implement the importance sampling approach for a bivariate local level model with time-varying level correlations.

We keep making the assumption of a stochastic time-varying state correlation. In Section 4.3 we discussed how to rewrite a bivariate local level model with stochastic time-varying level correlation, in such a way that the model becomes nonlinear Gaussian in the observation equation, and linear Gaussian in the state equation. The resulting model takes the form

$$\begin{aligned} \boldsymbol{z}_t &= Z_t(\boldsymbol{\theta}_t) + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{H}) \\ \boldsymbol{\theta}_t &= \boldsymbol{Z} \boldsymbol{\alpha}_t^* \\ \boldsymbol{\alpha}_{t+1}^* &= \boldsymbol{T} \alpha_t^* + \boldsymbol{\eta}_t^*, \quad \boldsymbol{\eta}_t^* \sim N(\mathbf{0}, \boldsymbol{Q}^*), \end{aligned} \tag{4.4.1}$$

for $t = 1, \ldots, T$, where $\boldsymbol{z}_t = (x_t, y_t)'$, $\boldsymbol{\alpha}_t^* = (\alpha_t^{y*}, \alpha_t^{x*}, \gamma_t)'$, $\boldsymbol{Z} = \boldsymbol{T} = \boldsymbol{I}_3$, $\boldsymbol{Q}^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma_\gamma^2 \end{bmatrix}$, and

$$Z_t(\boldsymbol{\theta}_t) = Z_t(\boldsymbol{\alpha}_t^*) = \left(\begin{array}{c} \sigma_y \alpha_t^{y*} \\ \sigma_x \left(\tanh(\gamma_y)\alpha_t^{y*} + \sqrt{1 - \tanh^2(\gamma_t)}\alpha_t^{x*} \right) \end{array} \right).$$

In model (4.4.1), the diagonal elements of $\boldsymbol{H}$ (σ_y and σ_x), and σ_γ are the parameters that need to be estimated by maximum likelihood. The off-diagonal elements of $\boldsymbol{H}$ are equal to zero.

The conditional log-density of $\boldsymbol{z}_t$ on $\boldsymbol{\theta}_t$ for model (4.4.1) is

$$\begin{aligned} \log p(\boldsymbol{z}_t|\boldsymbol{\theta}_t; \boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log(\det \boldsymbol{H}^{-1}) \\ - \frac{1}{2}(\boldsymbol{z}_t - Z_t(\boldsymbol{\theta}_t))'\boldsymbol{H}^{-1}(\boldsymbol{z}_t - Z_t(\boldsymbol{\theta}_t)), \end{aligned} \tag{4.4.2}$$

for $t = 1, \ldots, T$. Since the unobserved components of the nonlinear model (4.4.1) cannot be estimated by the standard Kalman filter (due to the nonlinearity of the model), the idea behind importance sampling is to first obtain a Gaussian linear state space model which approximates the nonlinear model (4.4.1) as much as possible. This approximate linear Gaussian model takes the following form

$$\begin{aligned} \boldsymbol{b}_t &= \boldsymbol{\theta}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(0, \boldsymbol{A}_t) \\ \boldsymbol{\theta}_t &= \boldsymbol{Z}\boldsymbol{\alpha}_t^* \\ \boldsymbol{\alpha}_{t+1}^* &= \boldsymbol{T}\boldsymbol{\alpha}_t^* + \boldsymbol{\eta}_t^*, \quad \boldsymbol{\eta}_t^* \sim N(0, \boldsymbol{Q}^*), \end{aligned} \tag{4.4.3}$$

for $t = 1, \ldots, T$. Notice that model (4.4.3) is now linear in the signal vector, $\boldsymbol{\theta}_t$, and the distribution of the error terms is still Gaussian. Moreover, the transition equations of models (4.4.1) and (4.4.3) are the same (i.e., the unconditional density of $\boldsymbol{\theta}_t$ is linear Gaussian in both models). The conditional log-density of $\boldsymbol{b}_t$ on $\boldsymbol{\theta}_t$ for model (4.4.3) is based on a linear Gaussian distribution, which we indicate with g, and takes the form

$$\log g(\boldsymbol{b}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta}) = -\frac{\dim(\boldsymbol{b}_t)}{2}\log(2\pi) \\ + \frac{1}{2}\log(\det \boldsymbol{A}_t^{-1}) - \frac{1}{2}(\boldsymbol{b}_t - \boldsymbol{\theta}_t)'\boldsymbol{A}_t^{-1}(\boldsymbol{b}_t - \boldsymbol{\theta}_t),$$

for $t = 1,\ldots,T$, where $\dim(\boldsymbol{b}_t)$ represents the dimension of the vector $\boldsymbol{b}_t$. Notice also that the observed vector, $\boldsymbol{b}_t$, and the covariance matrix of the innovations in the measurement equation, $\boldsymbol{A}_t$, in model (4.4.3), are different from, respectively, $\boldsymbol{z}_t$ and $\boldsymbol{H}$ in model (4.4.1). If these two new elements (which are called importance parameters) are chosen such that $\boldsymbol{A}_t = -\left[\ddot{p}(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})|_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t}\right]^{-1}$ and $\boldsymbol{b}_t = \hat{\boldsymbol{\theta}}_t + \boldsymbol{A}_t\,\dot{p}(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})|_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t}$, where $\dot{p}(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta}) = \frac{\partial \log p(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})}{\partial\boldsymbol{\theta}_t}$ and $\ddot{p}(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta}) = \frac{\partial^2 \log p(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})}{\partial\boldsymbol{\theta}_t\partial\boldsymbol{\theta}_t'}$, and with $\hat{\boldsymbol{\theta}}_t = \arg\max_{\boldsymbol{\theta}_t} p(\boldsymbol{\theta}_t|\boldsymbol{z}_t;\boldsymbol{\beta})$, which implies that $\hat{\boldsymbol{\theta}}_t$ is the mode of $p(\boldsymbol{\theta}_t|\boldsymbol{z}_t;\boldsymbol{\beta})$, then

$$\left.\frac{\partial \log g(\boldsymbol{b}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})}{\partial\boldsymbol{\theta}_t}\right|_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t} = \left.\frac{\partial \log p(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})}{\partial\boldsymbol{\theta}_t}\right|_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t}$$

$$\left.\frac{\partial^2 \log g(\boldsymbol{b}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})}{\partial\boldsymbol{\theta}_t\partial\boldsymbol{\theta}_t'}\right|_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t} = \left.\frac{\partial^2 \log p(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})}{\partial\boldsymbol{\theta}_t\partial\boldsymbol{\theta}_t'}\right|_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t},$$

for $t = 1,\ldots,T$. This last statement implies that the linear Gaussian model (4.4.3) approximates the nonlinear Gaussian model (4.4.1) up to the second order, since the first and second derivatives of the densities implied by both models, with respect to the signal vector, are the same at the mode $\hat{\boldsymbol{\theta}}_t$. In Appendix 4.C.1 we show how to derive the expressions for $\boldsymbol{A}_t$ and $\boldsymbol{b}_t$. Finally, Koopman et al. (2018) show that $g(\boldsymbol{b}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta}) = g(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})$, for $t = 1,\ldots,T$.

Shephard and Pitt (1997) and Durbin and Koopman (1997) explain that, given an initial guess for $\hat{\boldsymbol{\theta}}_t$ (for instance $\hat{\boldsymbol{\theta}}_t = \boldsymbol{0}$ for $t = 1,\ldots,T$), which yields initial guesses for $\boldsymbol{b}_t$ and $\boldsymbol{A}_t$ in model (4.4.3), it is possible to obtain the KFS estimate for $\boldsymbol{\theta}_t$, since model (4.4.3) is linear. The KFS estimate then replaces the initial guess for $\hat{\boldsymbol{\theta}}_t$, and a new KFS estimate for $\boldsymbol{\theta}_t$ can be obtained again for the new values of $\boldsymbol{b}_t$ and $\boldsymbol{A}_t$, and so on until convergence. The KFS estimate for $\boldsymbol{\theta}_t$ at convergence represents the mode $\hat{\boldsymbol{\theta}}_t$. The method just described for choosing $\boldsymbol{b}_t$ and $\boldsymbol{A}_t$, which is based on the second order approximation of the nonlinear model at the mode, is called SPDK (after Shephard and Pitt (1997) and Durbin and Koopman (1997)).

The necessity to find a Gaussian linear model that approximates the nonlinear one, is motivated by the fact that importance sampling aims at estimating $\mathrm{E}(\boldsymbol{\theta})$ by means of its Monte Carlo estimator (Durbin & Koopman, 2012, Chapter 11), which is expressed as

$$\hat{\boldsymbol{\theta}}_{\mathrm{M}} = \frac{1}{S}\sum_{i=1}^{S}\tilde{\boldsymbol{\theta}}^{(i)}, \text{ with draw } \tilde{\boldsymbol{\theta}}^{(i)} \sim p(\boldsymbol{\theta}|\boldsymbol{z};\boldsymbol{\beta}), \tag{4.4.4}$$

for S draws of $\tilde{\boldsymbol{\theta}}^{(i)}$ from $p(\boldsymbol{\theta}|\boldsymbol{z};\boldsymbol{\beta})$, where $\boldsymbol{\theta} = \boldsymbol{Z\alpha}$, with $\boldsymbol{\alpha}$ being the $mT \times 1$ state vector, and $\boldsymbol{z}$ the $nT \times 1$ observed vector. The problem is that it is not possible to draw from the conditional density $p(\boldsymbol{\theta}|\boldsymbol{z};\boldsymbol{\beta})$ implied in model (4.4.1), because of the nonlinearity of the model. Nonetheless, it is possible to rewrite (4.4.4) as

$$\hat{\boldsymbol{\theta}}_{\mathrm{M}} = \frac{1}{S}\sum_{i=1}^{S}\frac{\tilde{\boldsymbol{\theta}}^{(i)}p(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{z};\boldsymbol{\beta})}{g(\tilde{\boldsymbol{\theta}}^{(i)}|\boldsymbol{z};\boldsymbol{\beta})}, \text{ with draw } \tilde{\boldsymbol{\theta}}^{(i)} \sim g(\boldsymbol{\theta}|\boldsymbol{z};\boldsymbol{\beta}), \tag{4.4.5}$$

and draw $\tilde{\boldsymbol{\theta}}^{(i)}$ from the linear Gaussian conditional density $g(\boldsymbol{\theta}|\boldsymbol{z};\boldsymbol{\beta})$ implied by model (4.4.3), which is referred to as the importance density[2]. Since the transition equation of the nonlinear model (4.4.1) is linear Gaussian, it is possible to show (by simply applying the Bayes rule) that equation (4.4.5) boils

[2] Notice that the importance density is expressed in terms of $\boldsymbol{z}_t$ instead of $\boldsymbol{b}_t$ because, as mentioned earlier in this Section, Koopman et al. (2018) show that $g(\boldsymbol{b}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta}) = g(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})$, for $t = 1,\ldots,T$.

down to

$$\hat{\boldsymbol{\theta}}_{\mathrm{M}} = \frac{\sum_{i=1}^{S} \tilde{\boldsymbol{\theta}}^{(i)} w(\boldsymbol{z}|\tilde{\boldsymbol{\theta}}^{(i)};\boldsymbol{\beta})}{\sum_{i=1}^{S} w(\boldsymbol{z}|\tilde{\boldsymbol{\theta}}^{(i)};\boldsymbol{\beta})}, \text{ with draw } \tilde{\boldsymbol{\theta}}^{(i)} \sim g(\boldsymbol{\theta}|\boldsymbol{z};\boldsymbol{\beta}), \qquad (4.4.6)$$

where $w(\boldsymbol{z}|\tilde{\boldsymbol{\theta}}^{(i)};\boldsymbol{\beta}) = \frac{p(\boldsymbol{z}|\tilde{\boldsymbol{\theta}}^{(i)};\boldsymbol{\beta})}{g(\boldsymbol{z}|\tilde{\boldsymbol{\theta}}^{(i)};\boldsymbol{\beta})}$ are called importance weights.

Durbin and Koopman (2002) provide an algorithm for simulation smoothing (which holds only for linear Gaussian models), which allows us to draw $\tilde{\boldsymbol{\theta}}^{(i)}$ from $g(\boldsymbol{\theta}|\boldsymbol{z};\boldsymbol{\beta})$. The algorithm can be found in Appendix 4.C.2. For each draw $\tilde{\boldsymbol{\theta}}^{(i)}$, we evaluate $g(\boldsymbol{z}|\tilde{\boldsymbol{\theta}}^{(i)};\boldsymbol{\beta}) = \prod_{t=1}^{T} g(\boldsymbol{b}_t|\tilde{\boldsymbol{\theta}}_t^{(i)};\boldsymbol{\beta})$, with $\boldsymbol{b}_t|\tilde{\boldsymbol{\theta}}_t^{(i)} \sim N(\tilde{\boldsymbol{\theta}}_t^{(i)}, \boldsymbol{A}_t)$, and $p(\boldsymbol{z}|\tilde{\boldsymbol{\theta}}^{(i)};\boldsymbol{\beta}) = \prod_{t=1}^{T} p(\boldsymbol{z}_t|\tilde{\boldsymbol{\theta}}_t^{(i)};\boldsymbol{\beta})$, in order to compute the importance weights, and, finally, $\hat{\boldsymbol{\theta}}_M$. The hyperbolic tangent of the third and last element of $\hat{\boldsymbol{\theta}}_{t,M}$ represents the Monte Carlo estimator of the time-varying correlation at time t.

The parameter vector $\boldsymbol{\beta}$, which contains the static parameters of the model, is estimated by maximizing the following log-likelihood

$$\log \hat{p}(\boldsymbol{z};\boldsymbol{\beta}) = \log g(\boldsymbol{z};\boldsymbol{\beta}) + \log \left[\frac{1}{S} \sum_{i=1}^{S} w(\boldsymbol{z}|\tilde{\boldsymbol{\theta}}^{(i)};\boldsymbol{\beta}) \right],$$

where $\log g(\boldsymbol{z};\boldsymbol{\beta}) = \sum_{t=1}^{T} \log g(\boldsymbol{z}_t;\boldsymbol{\beta})$, with $\log g(\boldsymbol{z}_t;\boldsymbol{\beta})$ taking expression (4.2.1) and being evaluated via the Kalman filter recursions (4.2.2), with the difference that $\boldsymbol{z}_t = \boldsymbol{b}_t$ and $\boldsymbol{H} = \boldsymbol{A}_t$ evaluated at the mode $\hat{\boldsymbol{\theta}}_t$, for $t = 1, \ldots, T$.

4.5 Discussion and conclusions

We have just finished illustrating how the extended Kalman filter and importance sampling methods can attempt to deal with nonlinear state space models, where the nonlinearity is triggered by a stochastic state correlation. Both approaches require the linearisation of the model, and we did so by first bringing the nonlinearity from the transition equation to the observation equation. We therefore claimed that we could rewrite model (4.1.1) as

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{Z}\boldsymbol{C}_t\boldsymbol{\alpha}_t^* + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\alpha}_{t+1}^* &= \boldsymbol{T}\boldsymbol{\alpha}_t^* + \boldsymbol{\eta}_t^*, \quad \boldsymbol{\eta}_t^* \sim N(0, \boldsymbol{Q}^*), \quad t = 1, \dots, T. \end{aligned} \tag{4.5.1}$$

However, the attentive readers may have already realised that (4.5.1) is not a different way to rewrite model (4.1.1). Indeed, model (4.1.1) can be correctly rewritten as

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{Z}\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\alpha}_{t+1} &= \boldsymbol{T}\boldsymbol{\alpha}_t + \boldsymbol{C}_t\boldsymbol{\eta}_t^*, \quad \boldsymbol{\eta}_t^* \sim N(0, \boldsymbol{Q}^*), \quad t = 1, \dots, T. \end{aligned}$$

Now, since in the case of a bivariate local level model $\boldsymbol{T}$ is an identity matrix,

$$\boldsymbol{\alpha}_{t+1} = \sum_{i=1}^{t+1} \boldsymbol{C}_i\boldsymbol{\eta}_i^*, \quad \boldsymbol{\eta}_i^* \sim N(0, \boldsymbol{Q}^*), \quad t = 1, \dots, T,$$

and therefore, model (4.1.1) can, one last time, be correctly rewritten as

$$\begin{aligned} \boldsymbol{z}_t &= \boldsymbol{Z}\sum_{i=1}^{t} \boldsymbol{C}_i\boldsymbol{\eta}_i^* + \boldsymbol{\varepsilon}_t \\ \boldsymbol{\eta}_i^* &\sim N(0, \boldsymbol{Q}^*), \quad t = 1, \dots, T. \end{aligned}$$

In model (4.5.1) we were therefore missing a partial sum term which clearly renders the subsequent linearisation required by the extended Kalman filter and importance sampling methods, more cumbersome. These approaches have therefore not been investigated further. One possible, yet non-trivial, way out of this problem would be to employ the importance sampling approach by linearising the transition equation directly.

Although we could not show how to feasibly employ the two above-mentioned methods for the estimation of a stochastic state correlation, we illustrated that the estimation of a deterministic time-varying state correlation by the score-driven method, is feasible. However, it is not satisfactory, in terms of estimation accuracy of the correlation, with respect to the cubic splines method. This finding motivates the reason why we employed the latter approach, instead of the score-driven one, in order to create an approximate model (based on a deterministic time-varying specification for the correlation) for the indirect inference procedure of Chapter 3.

4.A GAS method: proofs and Monte Carlo simulation results

The following proofs apply to a more general state space model than the one considered in this chapter, where all matrices (i.e., $\boldsymbol{Z}_t$, $\boldsymbol{H}_t$, $\boldsymbol{T}_t$, $\boldsymbol{R}_t$ and $\boldsymbol{Q}_t$) are allowed to have time-varying parameters.

4.A.1 Derivation of equations (4.2.5)

$$
\begin{aligned}
\nabla_t = \frac{\partial \ell_t}{\partial \boldsymbol{\rho}_t} = \left[\frac{\partial \ell_t}{\partial \boldsymbol{\rho}_t'}\right]' &= \left[-\frac{1}{2}\frac{\partial \log\left(\det \boldsymbol{F}_t\right)}{\partial \boldsymbol{\rho}_t'} - \frac{1}{2}\frac{\partial \boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t}{\partial \boldsymbol{\rho}_t'}\right]' \\
&= -\frac{1}{2}\left[\frac{\partial \log\left(\det \boldsymbol{F}_t\right)}{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)'}\frac{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t}{\partial \boldsymbol{v}_t}\frac{\boldsymbol{v}_t}{\partial \boldsymbol{\rho}_t'}\right. \\
&\quad \left. + \frac{\partial \boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t}{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)'}\frac{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)}{\partial \boldsymbol{\rho}_t'}\right]' \\
&= -\frac{1}{2}\left[\operatorname{vec}\left(\boldsymbol{F}_t^{-1}\right)'\dot{\boldsymbol{F}}_t + \boldsymbol{v}_t'\left(\boldsymbol{F}_t^{-1} + \boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{V}}_t - \left(\boldsymbol{F}_t^{-1}\boldsymbol{v}_t \otimes \boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right)'\dot{\boldsymbol{F}}_t\right]' \\
&= -\frac{1}{2}\left[\dot{\boldsymbol{F}}_t'\operatorname{vec}\left(\boldsymbol{F}_t^{-1}\right) + 2\dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t - \dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\left(\boldsymbol{v}_t \otimes \boldsymbol{v}_t\right)\right] \\
&= -\frac{1}{2}\left[\dot{\boldsymbol{F}}_t'\operatorname{vec}\left(\boldsymbol{F}_t^{-1}\boldsymbol{F}_t\boldsymbol{F}_t^{-1}\right) - \dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\left(\boldsymbol{v}_t \otimes \boldsymbol{v}_t\right) + 2\dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right] \\
&= -\frac{1}{2}\left[\dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\operatorname{vec}\left(\boldsymbol{F}_t\right) - \dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\left(\boldsymbol{v}_t \otimes \boldsymbol{v}_t\right)\right. \\
&\quad \left. + 2\dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right] \\
&= -\frac{1}{2}\left[\dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\left[\operatorname{vec}\left(\boldsymbol{F}_t\right) - \left(\boldsymbol{v}_t \otimes \boldsymbol{v}_t\right)\right] + 2\dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right] \\
&= \frac{1}{2}\left[\dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\left[\left(\boldsymbol{v}_t \otimes \boldsymbol{v}_t\right) - \operatorname{vec}\left(\boldsymbol{F}_t\right)\right] - 2\dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right] \\
&= \frac{1}{2}\left[\dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\left[\operatorname{vec}\left(\boldsymbol{v}_t\boldsymbol{v}_t'\right) - \operatorname{vec}\left(\boldsymbol{F}_t\right)\right] - 2\dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right] \\
&= \frac{1}{2}\left[\dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\operatorname{vec}\left(\boldsymbol{v}_t\boldsymbol{v}_t' - \boldsymbol{F}_t\right) - 2\dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right].
\end{aligned}
\tag{4.A.1}
$$

$$
\begin{aligned}
\frac{\partial^2 \ell_t}{\partial \boldsymbol{\rho}_t \partial \boldsymbol{\rho}_t'} &= \frac{\partial \nabla_t}{\partial \boldsymbol{\rho}_t'} \\
&= \frac{\partial \nabla_t}{\partial \operatorname{vec}\left(\dot{\boldsymbol{F}}_t\right)'} \frac{\partial \operatorname{vec}\left(\dot{\boldsymbol{F}}_t\right)}{\boldsymbol{\rho}_t'} + \frac{\partial \nabla_t}{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)'} \frac{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)}{\boldsymbol{\rho}_t'} \\
&\quad + \frac{\partial \nabla_t}{\partial \operatorname{vec}\left(\dot{\boldsymbol{V}}_t\right)'} \frac{\partial \operatorname{vec}\left(\dot{\boldsymbol{V}}_t\right)}{\boldsymbol{\rho}_t'} + \frac{\partial \nabla_t}{\partial \boldsymbol{v}_t'} \frac{\partial \boldsymbol{v}_t}{\boldsymbol{\rho}_t'} \\
&= \frac{1}{2} \left[\operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right)' \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \otimes \boldsymbol{I}_{n^2}\right] \boldsymbol{C}_{n^2 k} \ddot{\boldsymbol{F}}_t \\
&\quad + \frac{1}{2} \left[\frac{\partial \dot{\boldsymbol{F}}_t \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right)}{\partial \operatorname{vec}\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)'} \frac{\partial \operatorname{vec}\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)}{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)'} \right. \\
&\quad \left. - \dot{\boldsymbol{F}}_t' \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)\right] \dot{\boldsymbol{F}}_t \\
&\quad + \left(\boldsymbol{v}_t \otimes \dot{\boldsymbol{V}}_t\right)' \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \dot{\boldsymbol{F}}_t - \left(\boldsymbol{v}_t' \boldsymbol{F}_t^{-1} \otimes \boldsymbol{I}_n\right) \boldsymbol{C}_{nk} \ddot{\boldsymbol{V}}_t \\
&\quad + \frac{1}{2} \dot{\boldsymbol{F}}_t' \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \left(\boldsymbol{v}_t \otimes \boldsymbol{I}_n + \boldsymbol{I}_n \otimes \boldsymbol{v}_t\right) \dot{\boldsymbol{V}}_t - \dot{\boldsymbol{V}}_t' \boldsymbol{F}_t^{-1} \dot{\boldsymbol{V}}_t,
\end{aligned} \tag{4.A.2}
$$

since

$$
\begin{aligned}
\frac{\partial \nabla_t}{\partial \operatorname{vec}\left(\dot{\boldsymbol{F}}_t\right)'} \frac{\partial \operatorname{vec}\left(\dot{\boldsymbol{F}}_t\right)}{\partial \boldsymbol{\rho}_t'} &= \frac{\partial \frac{1}{2} \dot{\boldsymbol{F}}_t' \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right) - \dot{\boldsymbol{V}}_t' \dot{\boldsymbol{F}}_t^{-1} \boldsymbol{v}_t}{\partial \operatorname{vec}\left(\dot{\boldsymbol{F}}_t\right)'} \\
&\quad \frac{\partial \operatorname{vec}\left(\dot{\boldsymbol{F}}_t\right)}{\partial \boldsymbol{\rho}_t'} \\
&= \frac{1}{2} \left[\operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right)' \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \otimes \boldsymbol{I}_{n^2}\right] \boldsymbol{C}_{n^2 k} \ddot{\boldsymbol{F}}_t,
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \nabla_t}{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)'} \frac{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)}{\partial \boldsymbol{\rho}_t'} &= \frac{\partial \frac{1}{2} \dot{\boldsymbol{F}}_t \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right) - \dot{\boldsymbol{V}}_t' \boldsymbol{F}_t^{-1} \boldsymbol{v}_t}{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)'} \\
&\quad \frac{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)}{\partial \boldsymbol{\rho}_t'} \\
&= \frac{1}{2} \left[\frac{\partial \dot{\boldsymbol{F}}_t \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right)}{\partial \operatorname{vec}\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)'} \right. \\
&\quad \frac{\partial \operatorname{vec}\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)}{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)'} \\
&\quad + \frac{\partial \dot{\boldsymbol{F}}_t \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right)}{\operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right)'} \\
&\quad \left. \frac{\partial \operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right)}{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)'} \right] \dot{\boldsymbol{F}}_t \\
&\quad + \left(\boldsymbol{F}_t^{-1} \boldsymbol{v}_t \otimes \boldsymbol{F}_t^{-1} \dot{\boldsymbol{V}}_t\right)' \dot{\boldsymbol{F}}_t \\
&= \frac{1}{2} \left[\frac{\partial \dot{\boldsymbol{F}}_t \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right)}{\partial \operatorname{vec}\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)'} \right. \\
&\quad \left. \frac{\partial \operatorname{vec}\left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right)}{\partial \operatorname{vec}\left(\boldsymbol{F}_t\right)'} - \dot{\boldsymbol{F}}_t' \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \right] \dot{\boldsymbol{F}}_t \\
&\quad + \left(\boldsymbol{v}_t \otimes \dot{\boldsymbol{V}}_t\right)' \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \dot{\boldsymbol{F}}_t,
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \nabla_t}{\partial \operatorname{vec}\left(\dot{\boldsymbol{V}}_t\right)'} \frac{\partial \operatorname{vec}\left(\dot{\boldsymbol{V}}_t\right)}{\partial \boldsymbol{\rho}_t'} &= \frac{\partial \frac{1}{2} \dot{\boldsymbol{F}}_t \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{F}_t^{-1}\right) \operatorname{vec}\left(\boldsymbol{v}_t \boldsymbol{v}_t' - \boldsymbol{F}_t\right) - \dot{\boldsymbol{V}}_t' \dot{\boldsymbol{F}}_t^{-1} \boldsymbol{v}_t}{\partial \operatorname{vec}\left(\dot{\boldsymbol{V}}_t\right)'} \\
&\quad \frac{\partial \operatorname{vec}\left(\dot{\boldsymbol{V}}_t\right)}{\partial \boldsymbol{\rho}_t'} \\
&= -\left(\boldsymbol{v}_t' \boldsymbol{F}_t^{-1} \otimes \boldsymbol{I}_n\right) \boldsymbol{C}_{nk} \ddot{\boldsymbol{V}}_t,
\end{aligned}
$$

$$\frac{\partial \nabla_t}{\partial \boldsymbol{v}_t'}\frac{\partial \boldsymbol{v}_t}{\partial \boldsymbol{\rho}_t'} = \frac{\partial \frac{1}{2}\dot{\boldsymbol{F}}_t\left(\boldsymbol{F}_t^{-1}\otimes \boldsymbol{F}_t^{-1}\right)\operatorname{vec}\left(\boldsymbol{v}_t\boldsymbol{v}_t' - \boldsymbol{F}_t\right) - \dot{\boldsymbol{V}}_t'\dot{\boldsymbol{F}}_t^{-1}\boldsymbol{v}_t}{\partial \boldsymbol{v}_t'}\frac{\partial \boldsymbol{v}_t}{\partial \boldsymbol{\rho}_t'}$$
$$= \frac{1}{2}\dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1}\otimes \boldsymbol{F}_t^{-1}\right)\left(\boldsymbol{v}_t\otimes \boldsymbol{I}_n + \boldsymbol{I}_n\otimes \boldsymbol{v}_t\right)\dot{\boldsymbol{V}}_t - \dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\dot{\boldsymbol{V}}_t.$$

$\ddot{\boldsymbol{F}}_t = \frac{\partial \operatorname{vec}\left(\dot{\boldsymbol{F}}_t\right)}{\partial \boldsymbol{\rho}_t'}$ and $\ddot{\boldsymbol{V}}_t = \frac{\partial \operatorname{vec}\left(\dot{\boldsymbol{V}}_t\right)}{\partial \boldsymbol{\rho}_t'}$.

$$\boldsymbol{\mathcal{I}}_t = -\operatorname{E}_{t-1}\left(\frac{\partial^2 \ell_t}{\partial \boldsymbol{\rho}_t\partial \boldsymbol{\rho}_t'}\right) = \frac{1}{2}\dot{\boldsymbol{F}}_t'\left(\boldsymbol{F}_t^{-1}\otimes \boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{F}}_t + \dot{\boldsymbol{V}}_t'\boldsymbol{F}_t^{-1}\dot{\boldsymbol{V}}_t, \quad (4.A.3)$$

since $\boldsymbol{v}_t$ is the only random element in equation (4.A.2), $\operatorname{E}_{t-1}(\boldsymbol{v}_t) = \mathbf{0}$ and $\operatorname{E}_{t-1}(\boldsymbol{v}_t\boldsymbol{v}_t') = \boldsymbol{F}_t$. $\operatorname{E}_{t-1}\left(\dot{\boldsymbol{V}}_t\right) = \operatorname{E}_{t-1}\left(-\boldsymbol{Z}_t\dot{\boldsymbol{A}}_t\right) = -\boldsymbol{Z}_t\dot{\boldsymbol{A}}_t$ since $\dot{\boldsymbol{A}}_t$ depends on past information which is known at time t. The same argument holds for $\ddot{\boldsymbol{V}}_t$. $\dot{\boldsymbol{V}}_t$ and $\ddot{\boldsymbol{V}}_t$ are therefore non-random given the conditional expectation at time $t-1$. Moreover, $\operatorname{E}_{t-1}(\nabla_t) = \mathbf{0}$.

4.A.2 Derivation of the additional Kalman filter recursions (4.2.6)

$$\underset{n\times k}{\dot{\boldsymbol{V}}_t} = \frac{\partial \boldsymbol{v}_t}{\partial \boldsymbol{\rho}_t'}$$
$$= \frac{\partial \boldsymbol{v}_t}{\partial \operatorname{vec}(\boldsymbol{Z}_t)'}\frac{\partial \operatorname{vec}(\boldsymbol{Z}_t)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \boldsymbol{v}_t}{\partial \boldsymbol{a}_t'}\frac{\partial \boldsymbol{a}_t}{\partial \boldsymbol{\rho}_t'}$$
$$= -\left[\left(\boldsymbol{a}_t'\otimes \boldsymbol{I}_n\right)\dot{\boldsymbol{Z}}_t + \boldsymbol{Z}_t\dot{\boldsymbol{A}}_t\right]$$
$$= -\boldsymbol{Z}\dot{\boldsymbol{A}}_t.$$

$$
\begin{aligned}
\underset{n^2\times k}{\dot{\boldsymbol{F}}_t} &= \frac{\partial\,\mathrm{vec}\left(\boldsymbol{F}_t\right)}{\partial\boldsymbol{\rho}_t'} \\
&= \frac{\partial\,\mathrm{vec}\left(\boldsymbol{F}_t\right)}{\partial\,\mathrm{vec}\left(\boldsymbol{Z}_t\right)'}\frac{\partial\,\mathrm{vec}\left(\boldsymbol{Z}_t\right)}{\partial\boldsymbol{\rho}_t'} + \frac{\partial\,\mathrm{vec}\left(\boldsymbol{F}_t\right)}{\partial\,\mathrm{vec}\left(\boldsymbol{P}_t\right)'}\frac{\partial\,\mathrm{vec}\left(\boldsymbol{P}_t\right)}{\partial\boldsymbol{\rho}_t'} \\
&\quad + \frac{\partial\,\mathrm{vec}\left(\boldsymbol{F}_t\right)}{\partial\,\mathrm{vec}\left(\boldsymbol{H}_t\right)'}\frac{\partial\,\mathrm{vec}\left(\boldsymbol{H}_t\right)}{\partial\boldsymbol{\rho}_t'} \\
&= \left(\boldsymbol{I}_{n^2} + \boldsymbol{C}_n\right)\left(\boldsymbol{Z}_t\boldsymbol{P}_t \otimes \boldsymbol{I}_n\right)\dot{\boldsymbol{Z}}_t + \left(\boldsymbol{Z}_t \otimes \boldsymbol{Z}_t\right)\dot{\boldsymbol{P}}_t + \dot{\boldsymbol{H}}_t \\
&= \left(\boldsymbol{Z} \otimes \boldsymbol{Z}\right)\dot{\boldsymbol{P}}_t.
\end{aligned}
$$

$$
\begin{aligned}
\underset{mn\times k}{\dot{\boldsymbol{K}}_t} &= \frac{\partial\,\mathrm{vec}\left(\boldsymbol{K}_t\right)}{\partial\boldsymbol{\rho}_t'} \\
&= \frac{\partial\,\mathrm{vec}\left(\boldsymbol{K}_t\right)}{\partial\,\mathrm{vec}\left(\boldsymbol{T}_t\right)'}\frac{\partial\,\mathrm{vec}\left(\boldsymbol{T}_t\right)}{\partial\boldsymbol{\rho}_t'} + \frac{\partial\,\mathrm{vec}\left(\boldsymbol{K}_t\right)}{\partial\,\mathrm{vec}\left(\boldsymbol{P}_t\right)'}\frac{\partial\,\mathrm{vec}\left(\boldsymbol{P}_t\right)}{\partial\boldsymbol{\rho}_t'} \\
&\quad + \frac{\partial\,\mathrm{vec}\left(\boldsymbol{K}_t\right)}{\partial\,\mathrm{vec}\left(\boldsymbol{Z}_t\right)'}\frac{\partial\,\mathrm{vec}\left(\boldsymbol{Z}_t\right)}{\partial\boldsymbol{\rho}_t'} + \frac{\partial\,\mathrm{vec}\left(\boldsymbol{K}_t\right)}{\partial\,\mathrm{vec}\left(\boldsymbol{F}_t\right)'}\frac{\partial\,\mathrm{vec}\left(\boldsymbol{F}_t\right)}{\partial\boldsymbol{\rho}_t'} \\
&= \left(\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t\boldsymbol{P}_t \otimes \boldsymbol{I}_m\right)\dot{\boldsymbol{T}}_t + \left(\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t \otimes \boldsymbol{T}_t\right)\dot{\boldsymbol{P}}_t + \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{T}_t\boldsymbol{P}_t\right)\boldsymbol{C}_{nm}\dot{\boldsymbol{Z}}_t \\
&\quad - \left(\boldsymbol{I}_n\boldsymbol{F}_t^{-1} \otimes \boldsymbol{T}_t\boldsymbol{P}_t\boldsymbol{Z}_t'\boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{F}}_t \\
&= \left(\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t\boldsymbol{P}_t \otimes \boldsymbol{I}_m\right)\dot{\boldsymbol{T}}_t + \left(\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t \otimes \boldsymbol{T}_t\right)\dot{\boldsymbol{P}}_t + \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{T}_t\boldsymbol{P}_t\right)\boldsymbol{C}_{nm}\dot{\boldsymbol{Z}}_t \\
&\quad - \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{K}_t\right)\dot{\boldsymbol{F}}_t \\
&= \left(\boldsymbol{F}_t^{-1}\boldsymbol{Z} \otimes \boldsymbol{T}\right)\dot{\boldsymbol{P}}_t - \left(\boldsymbol{F}_t^{-1} \otimes \boldsymbol{K}_t\right)\dot{\boldsymbol{F}}_t.
\end{aligned}
$$

$$
\begin{aligned}
\underset{m\times k}{\dot{\boldsymbol{A}}_{t|t}} &= \frac{\partial \boldsymbol{a}_{t|t}}{\partial \boldsymbol{\rho}_t'} \\
&= \frac{\partial \boldsymbol{a}_{t|t}}{\partial \boldsymbol{a}_t'} \frac{\partial \boldsymbol{a}_t}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \boldsymbol{a}_{t|t}}{\partial \operatorname{vec}(\boldsymbol{P}_t)'} \frac{\partial \operatorname{vec}(\boldsymbol{P}_t)}{\partial \boldsymbol{\rho}_t'} \\
&\quad + \frac{\partial \boldsymbol{a}_{t|t}}{\partial \operatorname{vec}(\boldsymbol{Z}_t)'} \frac{\partial \operatorname{vec}(\boldsymbol{Z}_t)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \boldsymbol{a}_{t|t}}{\partial \operatorname{vec}(\boldsymbol{F}_t)'} \frac{\partial \operatorname{vec}(\boldsymbol{F}_t)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \boldsymbol{a}_{t|t}}{\partial \boldsymbol{v}_t'} \frac{\partial \boldsymbol{v}_t}{\partial \boldsymbol{\rho}_t'} \\
&= \dot{\boldsymbol{A}}_t + \left(\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t \otimes \boldsymbol{I}_m\right)\dot{\boldsymbol{P}}_t + \left(\boldsymbol{v}_t'\boldsymbol{F}_t^{-1} \otimes \boldsymbol{P}_t\right)\boldsymbol{C}_{nm}\dot{\boldsymbol{Z}}_t \\
&\quad - \left(\boldsymbol{v}_t'\boldsymbol{F}_t^{-1} \otimes \boldsymbol{P}_t\boldsymbol{Z}_t'\boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{F}}_t + \boldsymbol{P}_t\boldsymbol{Z}_t'\boldsymbol{F}_t^{-1}\dot{\boldsymbol{V}}_t \\
&= \dot{\boldsymbol{A}}_t + \left(\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{Z} \otimes \boldsymbol{I}_m\right)\dot{\boldsymbol{P}}_t - \left(\boldsymbol{v}_t'\boldsymbol{F}_t^{-1} \otimes \boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{F}}_t + \boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\dot{\boldsymbol{V}}_t.
\end{aligned}
$$

$$
\begin{aligned}
\underset{m^2\times k}{\dot{\boldsymbol{P}}_{t|t}} &= \frac{\partial \operatorname{vec}\left(\boldsymbol{P}_{t|t}\right)}{\partial \boldsymbol{\rho}_t'} \\
&= \frac{\partial \operatorname{vec}\left(\boldsymbol{P}_{t|t}\right)}{\partial \operatorname{vec}(\boldsymbol{P}_t)'} \frac{\partial \operatorname{vec}(\boldsymbol{P}_t)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \operatorname{vec}\left(\boldsymbol{P}_{t|t}\right)}{\partial \operatorname{vec}(\boldsymbol{Z}_t)'} \frac{\partial \operatorname{vec}(\boldsymbol{Z}_t)}{\partial \boldsymbol{\rho}_t'} \\
&\quad + \frac{\partial \operatorname{vec}\left(\boldsymbol{P}_{t|t}\right)}{\partial \operatorname{vec}(\boldsymbol{F}_t)'} \frac{\partial \operatorname{vec}(\boldsymbol{F}_t)}{\partial \boldsymbol{\rho}_t'} \\
&= \dot{\boldsymbol{P}}_t - \left[\boldsymbol{P}_t\boldsymbol{Z}_t'\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t \otimes \boldsymbol{I}_m + \boldsymbol{I}_m \otimes \boldsymbol{P}_t\boldsymbol{Z}_t'\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t\right]\dot{\boldsymbol{P}}_t \\
&\quad - \left(\boldsymbol{P}_t \otimes \boldsymbol{P}_t\right)\left(\boldsymbol{I}_{n^2} + \boldsymbol{C}_n\right)\left(\boldsymbol{Z}_t\boldsymbol{F}_t^{-1} \otimes \boldsymbol{I}_n\right)\dot{\boldsymbol{Z}}_t \\
&\quad + \left(\boldsymbol{P}_t\boldsymbol{Z}_t'\boldsymbol{F}_t^{-1} \otimes \boldsymbol{P}_t\boldsymbol{Z}_t'\boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{F}}_t \\
&= \dot{\boldsymbol{P}}_t - \left[\boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\boldsymbol{Z} \otimes \boldsymbol{I}_m + \boldsymbol{I}_m \otimes \boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\boldsymbol{Z}\right]\dot{\boldsymbol{P}}_t \\
&\quad + \left(\boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1} \otimes \boldsymbol{P}_t\boldsymbol{Z}'\boldsymbol{F}_t^{-1}\right)\dot{\boldsymbol{F}}_t.
\end{aligned}
$$

$$\begin{aligned}
\underset{m\times k}{\dot{\boldsymbol{A}}_{t+1}} &= \frac{\partial \boldsymbol{a}_{t+1}}{\partial \boldsymbol{\rho}_t'} \\
&= \frac{\partial \boldsymbol{a}_{t+1}}{\partial \operatorname{vec}(\boldsymbol{T}_t)'}\frac{\partial \operatorname{vec}(\boldsymbol{T}_t)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \boldsymbol{a}_{t+1}}{\partial \boldsymbol{a}_t'}\frac{\partial \operatorname{vec} \boldsymbol{a}_t}{\partial \boldsymbol{\rho}_t'} \\
&\quad + \frac{\partial \boldsymbol{a}_{t+1}}{\partial \operatorname{vec}(\boldsymbol{K}_t)'}\frac{\partial \operatorname{vec}(\boldsymbol{K}_t)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \boldsymbol{a}_{t+1}}{\partial \boldsymbol{v}_t'}\frac{\partial \boldsymbol{v}_t}{\partial \boldsymbol{\rho}_t'} \\
&= \left(\boldsymbol{a}_t' \otimes \boldsymbol{I}_m\right)\dot{\boldsymbol{T}}_t + \boldsymbol{T}_t\dot{\boldsymbol{A}}_t + \left(\boldsymbol{v}_t' \otimes \boldsymbol{I}_m\right)\dot{\boldsymbol{K}}_t + \boldsymbol{K}_t\dot{\boldsymbol{V}}_t \\
&= \boldsymbol{T}\dot{\boldsymbol{A}}_t + \left(\boldsymbol{v}_t' \otimes \boldsymbol{I}_m\right)\dot{\boldsymbol{K}}_t + \boldsymbol{K}_t\dot{\boldsymbol{V}}_t.
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{P}_{t+1} &= \boldsymbol{T}_t\boldsymbol{P}_t\left(\boldsymbol{T}_t' - \boldsymbol{Z}_t'\boldsymbol{K}_t'\right) + \boldsymbol{R}_t\boldsymbol{Q}_t\boldsymbol{R}_t' \\
&= \boldsymbol{T}_t\boldsymbol{P}_t\boldsymbol{T}_t' - \boldsymbol{T}_t\boldsymbol{P}_t\boldsymbol{Z}_t'\boldsymbol{K}_t' + \boldsymbol{R}_t\boldsymbol{Q}_t\boldsymbol{R}_t'.
\end{aligned}$$

$$\begin{aligned}
\underset{m^2\times k}{\dot{\boldsymbol{P}}_{t+1}} &= \frac{\partial \operatorname{vec}(\boldsymbol{P}_{t+1})}{\partial \boldsymbol{\rho}_t'} \\
&= \frac{\partial \operatorname{vec}(\boldsymbol{P}_{t+1})}{\partial \operatorname{vec}(\boldsymbol{T}_t)'}\frac{\partial \operatorname{vec}(\boldsymbol{T}_t)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \operatorname{vec}(\boldsymbol{P}_{t+1})}{\partial \operatorname{vec}(\boldsymbol{P}_t)'}\frac{\partial \operatorname{vec}(\boldsymbol{P}_t)}{\partial \boldsymbol{\rho}_t'} \\
&\quad + \frac{\partial \operatorname{vec}(\boldsymbol{P}_{t+1})}{\partial \operatorname{vec}(\boldsymbol{K}_t)'}\frac{\partial \operatorname{vec}(\boldsymbol{K}_t)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \operatorname{vec}(\boldsymbol{P}_{t+1})}{\partial \operatorname{vec}(\boldsymbol{Z}_t)'}\frac{\partial \operatorname{vec}(\boldsymbol{Z}_t)}{\partial \boldsymbol{\rho}_t'} \\
&\quad + \frac{\partial \operatorname{vec}(\boldsymbol{P}_{t+1})}{\partial \operatorname{vec}(\boldsymbol{R}_t)'}\frac{\partial \operatorname{vec}(\boldsymbol{R}_t)}{\partial \boldsymbol{\rho}_t'} + \frac{\partial \operatorname{vec}(\boldsymbol{P}_{t+1})}{\partial \operatorname{vec}(\boldsymbol{Q}_t)'}\frac{\partial \operatorname{vec}(\boldsymbol{Q}_t)}{\partial \boldsymbol{\rho}_t'} \\
&= \left[\left(\boldsymbol{I}_{m^2} + \boldsymbol{C}_m\right)\left(\boldsymbol{T}_t\boldsymbol{P}_t \otimes \boldsymbol{I}_m\right) - \left(\boldsymbol{K}_t\boldsymbol{Z}_t\boldsymbol{P}_t \otimes \boldsymbol{I}_m\right)\right]\dot{\boldsymbol{T}}_t \\
&\quad + \left[\left(\boldsymbol{T}_t \otimes \boldsymbol{T}_t\right) - \left(\boldsymbol{K}_t\boldsymbol{Z}_t \otimes \boldsymbol{T}_t\right)\right]\dot{\boldsymbol{P}}_t \\
&\quad - \left(\boldsymbol{I}_m \otimes \boldsymbol{T}_t\boldsymbol{P}_t\boldsymbol{Z}_t'\right)\boldsymbol{C}_{mn}\dot{\boldsymbol{K}}_t - \left(\boldsymbol{K}_t \otimes \boldsymbol{T}_t\boldsymbol{P}_t\right)\boldsymbol{C}_{nm}\dot{\boldsymbol{Z}}_t \\
&\quad + \left(\boldsymbol{I}_{m^2} + \boldsymbol{C}_m\right)\left(\boldsymbol{R}_t\boldsymbol{Q}_t \otimes \boldsymbol{I}_m\right)\dot{\boldsymbol{R}}_t + \left(\boldsymbol{R}_t \otimes \boldsymbol{R}_t\right)\dot{\boldsymbol{Q}}_t \\
&= \left[\left(\boldsymbol{T} \otimes \boldsymbol{T}\right) - \left(\boldsymbol{K}_t\boldsymbol{Z} \otimes \boldsymbol{T}\right)\right]\dot{\boldsymbol{P}}_t - \left(\boldsymbol{I}_m \otimes \boldsymbol{T}\boldsymbol{P}_t\boldsymbol{Z}'\right)\boldsymbol{C}_{mn}\dot{\boldsymbol{K}}_t \\
&\qquad + \left(\boldsymbol{R} \otimes \boldsymbol{R}\right)\dot{\boldsymbol{Q}}_t.
\end{aligned}$$

All the last equalities hold if $\boldsymbol{Q}_t$ is the only matrix which contains time-varying parameters.

4.A.3 Monte Carlo simulation results

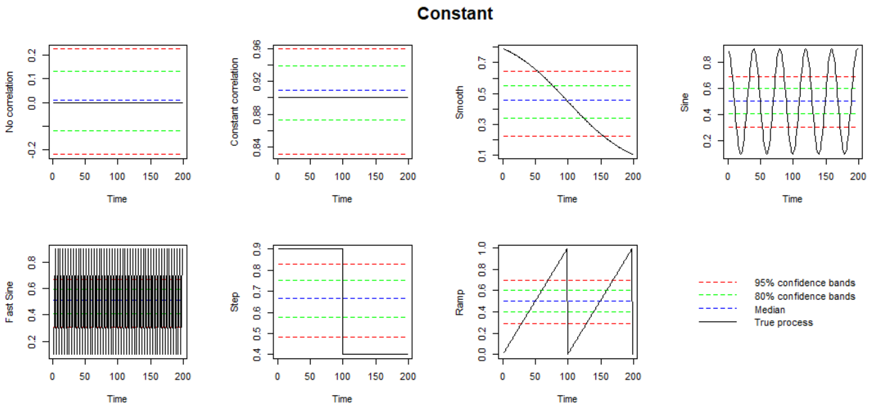


Figure 4.A.1: Bivariate local level model. $T = 200$, $n_{\text{sim}} = 500$.

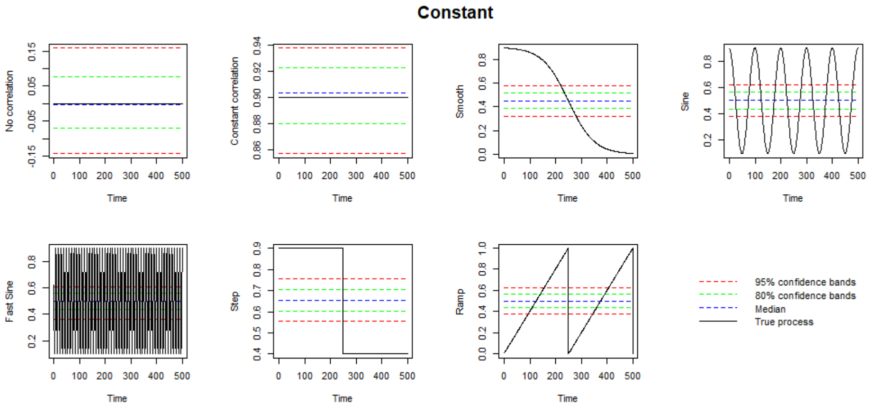


Figure 4.A.2: Bivariate local level model. $T = 500$, $n_{\text{sim}} = 500$.

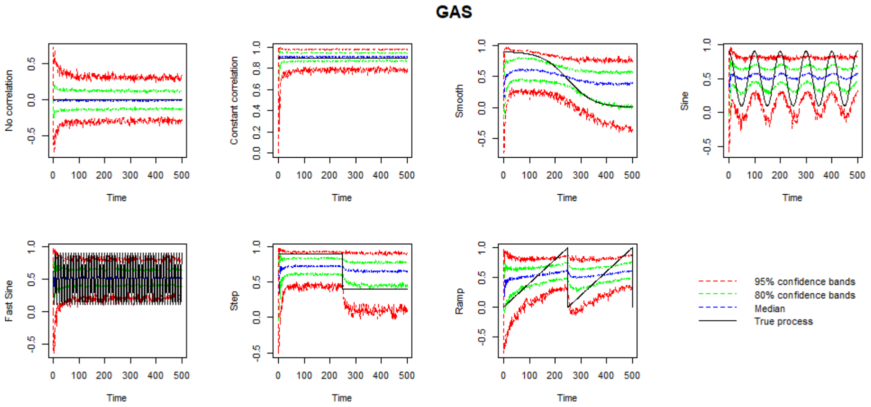


Figure 4.A.3: Bivariate local level model. $T = 500$, $n_{\text{sim}} = 500$.

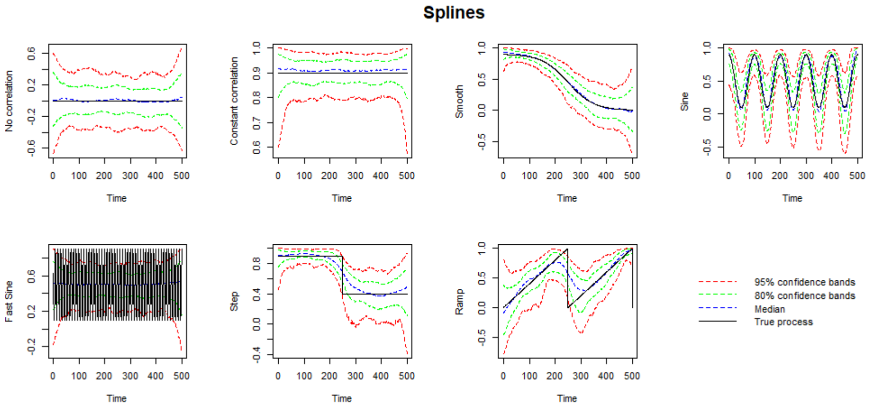


Figure 4.A.4: Bivariate local level model. $T = 500$, $n_{\text{sim}} = 500$.

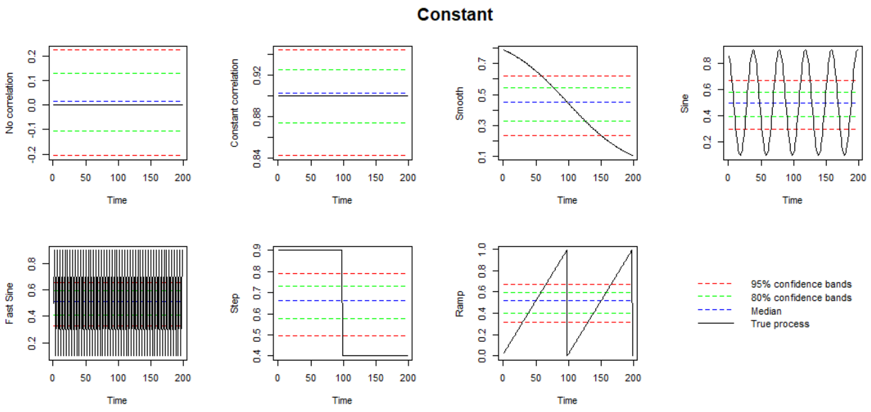


Figure 4.A.5: Bivariate smooth trend model. $T = 200$, $n_{\text{sim}} = 500$.

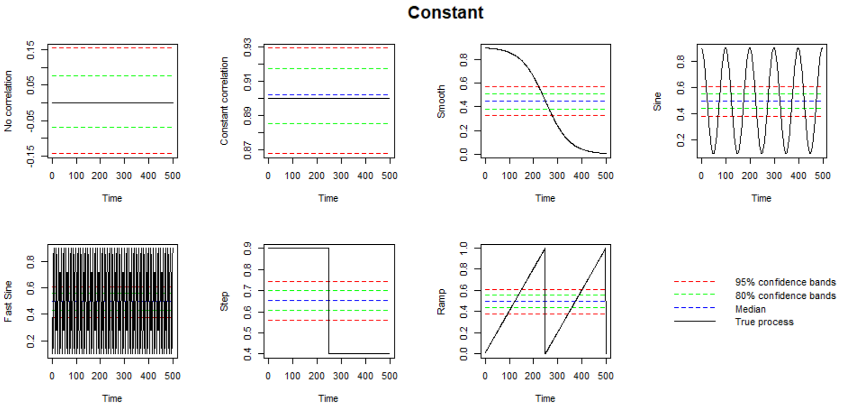


Figure 4.A.6: Bivariate smooth trend model. $T = 500$, $n_{\text{sim}} = 500$.

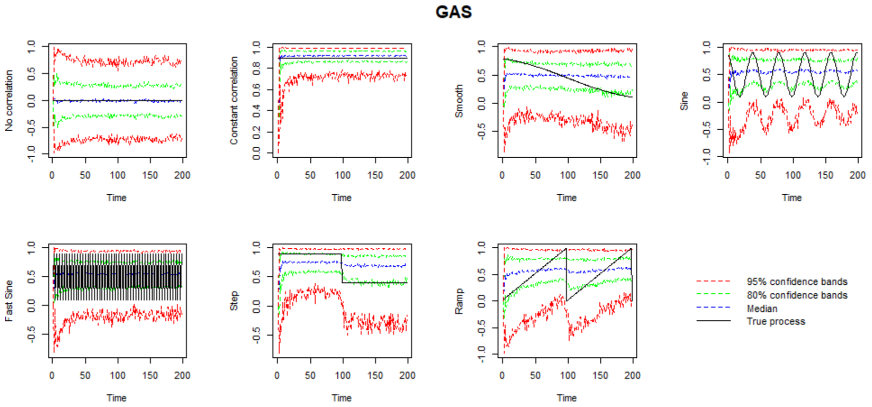


Figure 4.A.7: Bivariate smooth trend model. $T = 200$, $n_{\text{sim}} = 500$.

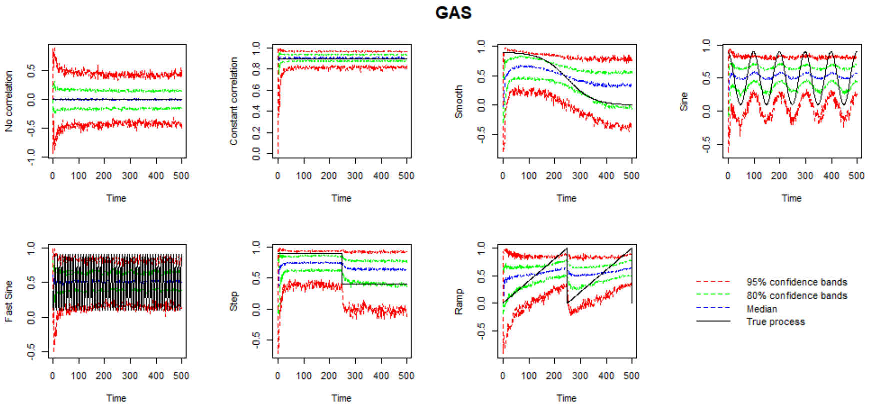


Figure 4.A.8: Bivariate smooth trend model. $T = 500$, $n_{\text{sim}} = 500$.

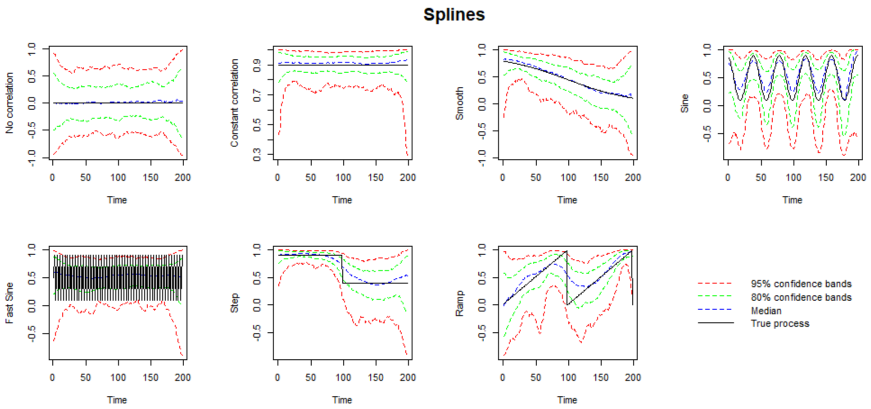


Figure 4.A.9: Bivariate smooth trend model. $T = 200$, $n_{\text{sim}} = 500$.

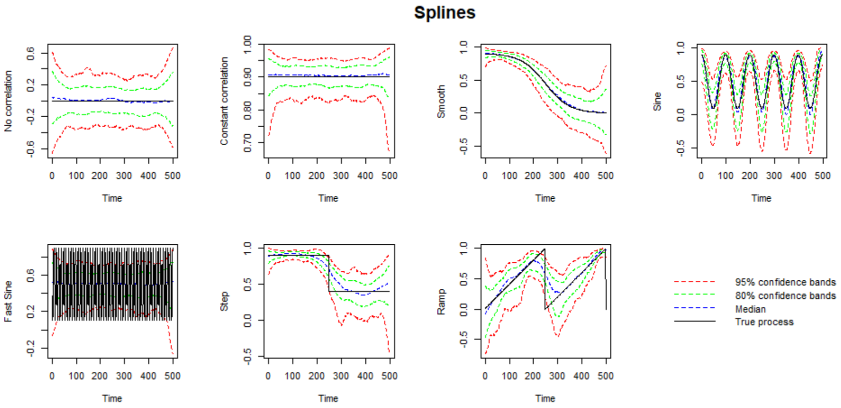


Figure 4.A.10: Bivariate smooth trend model. $T = 500$, $n_{\text{sim}} = 500$.

4.B Extended Kalman filter: bivariate smooth trend model

The observation equations for a bivariate smooth trend model are:

$$\begin{aligned} y_t &= L_t^y + \varepsilon_t^y \\ x_t &= L_t^x + \varepsilon_t^x, \quad t = 1, \ldots, T, \end{aligned} \tag{4.B.1}$$

with $(\varepsilon_t^y, \varepsilon_t^x)' \sim N(\mathbf{0}, \boldsymbol{H})$.

The transition equations are:

$$\begin{aligned} L_{t+1}^y &= L_t^y + R_t^y \\ R_{t+1}^y &= R_t^y + \eta_t^y \\ L_{t+1}^x &= L_t^x + R_t^x \\ R_{t+1}^x &= R_t^x + \eta_t^x \\ \gamma_{t+1} &= \gamma_t + \eta_t^\gamma, \quad t = 1, \ldots, T, \end{aligned}$$

with $(\eta_t^y, \eta_t^x, \eta_t^\gamma)' \sim N(\mathbf{0}, \boldsymbol{Q}_t)$, and $\boldsymbol{Q}_t = \begin{bmatrix} \sigma_y^2 & \tanh(\gamma_t)\sigma_y\sigma_x & 0 \\ \tanh(\gamma_t)\sigma_y\sigma_x & \sigma_x^2 & 0 \\ 0 & 0 & \sigma_\gamma^2 \end{bmatrix}$.
The states L_t and R_t represent, respectively, the level and the slope of the trend.

The bivariate smooth trend model can be written in compact notation as equation (4.1.1), with $\boldsymbol{z}_t = (x_t, y_t)'$, $\boldsymbol{\varepsilon}_t = (\varepsilon_t^y, \varepsilon_t^x)'$, $\boldsymbol{\alpha}_t = (L_t^y, R_t^y, L_t^x, R_t^x, \gamma_t)'$, $\boldsymbol{\eta}_t = (\eta_t^y, \eta_t^x, \eta_t^\gamma)'$, $\boldsymbol{Z} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$, $\boldsymbol{T} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$, and

$$\boldsymbol{R} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Similarly as seen for the local level model in Section 4.3, it is possible to transform the model in such a way that the nonlinearity involves only the state variables, and not the innovation terms, by means of the Cholesky decomposition of $\boldsymbol{Q}_t$. Again the nonlinearity has to be tackled in the measurement equations, which become:

$$\begin{aligned} y_t &= \sigma_y L_t^{y*} + \varepsilon_t^y \\ x_t &= \sigma_x \left(\tanh(\gamma_t) L_t^{y*} + \sqrt{1 - \tanh^2(\gamma_t)} L_t^{x*} \right) + \varepsilon_t^x, \quad t = 1, \ldots, T, \end{aligned} \tag{4.B.2}$$

and which have the same form as equation (4.3.1). The transition equations of the transformed model are:

$$\begin{aligned} L_{t+1}^{y*} &= L_t^{y*} + R_t^{y*} \\ R_{t+1}^{y*} &= R_t^{y*} + \eta_t^{y*} \\ L_{t+1}^{x*} &= L_t^{x*} + R_t^{x*} \\ R_{t+1}^{x*} &= R_t^{x*} + \eta_t^{x*} \\ \gamma_{t+1} &= \gamma_t + \eta_t^{\gamma}, \quad t = 1, \ldots, T, \end{aligned}$$

with $\left(\eta_t^{y*}, \eta_t^{x*}, \eta_t^{\gamma}\right)' \sim N\left(\mathbf{0}, \boldsymbol{Q}^*\right)$ and $\boldsymbol{Q}^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma_\gamma^2 \end{bmatrix}.$

The procedure for the linearisation of the model is the same as the one discussed in Section 4.3, with the difference that now

$$
\begin{aligned}
\dot{\boldsymbol{Z}}_t &= \left.\frac{\partial Z_t(\boldsymbol{\alpha}_t^*)}{\partial \boldsymbol{\alpha}_t^{*\prime}}\right|_{\boldsymbol{\alpha}_t^*=\boldsymbol{a}_t} \\
&= \left.\frac{\partial Z_t(\boldsymbol{\alpha}_t^*)}{\partial\left(L_t^{y*}, R_t^{y*}, L_t^{x*}, R_t^{x*}, \gamma_t\right)}\right|_{\boldsymbol{\alpha}_t^*=\boldsymbol{a}_t} \\
&= \left[\begin{array}{cccc} \sigma_y & 0 & 0 & 0 \\ \sigma_x \tanh(\gamma_t) & 0 & \sigma_x\sqrt{1-\tanh^2(\gamma_t)} & 0 \end{array}\right. \\
&\qquad \left.\left.\begin{array}{c} 0 \\ \sigma_x\sqrt{1-\tanh^2(\gamma_t)}\left(\sqrt{1-\tanh^2(\gamma_t)}\alpha_t^{y*} - \tanh(\gamma_t)\alpha_t^{x*}\right) \end{array}\right]\right|_{\boldsymbol{\alpha}_t^*=\boldsymbol{a}_t}.
\end{aligned}
$$

The extended recursions for the Kalman filter and smoother are the same as the ones discussed in Section 4.3.

4.C Importance sampling: importance parameters and simulation smoothing

4.C.1 Derivation of the expressions for the importance parameters

Given the conditional log-density (4.4.2) of the nonlinear Gaussian model (4.4.1),

$$
\dot{p}(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta}) = \frac{\partial \log p(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})}{\partial \boldsymbol{\theta}_t} = \dot{\boldsymbol{Z}}_t' \boldsymbol{H}^{-1}(\boldsymbol{z}_t - Z_t(\boldsymbol{\theta}_t)),
$$

where

$$\begin{aligned}
\dot{\boldsymbol{Z}}_t &= \frac{\partial Z_t(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t'} \\
&= \frac{\partial Z_t(\boldsymbol{\alpha}_t^*)}{\partial \boldsymbol{\alpha}_t^{*\prime}} \\
&= \frac{\partial Z_t(\boldsymbol{\alpha}_t^*)}{\partial \left(\alpha_t^{y*}, \alpha_t^{x*}, \gamma_t\right)} \\
&= \left[\begin{array}{cc} \sigma_y & 0 \\ \sigma_x \tanh(\gamma_t) & \sigma_x \sqrt{1-\tanh^2(\gamma_t)} \end{array} \right. \\
&\qquad \left. \begin{array}{c} 0 \\ \sigma_x \sqrt{1-\tanh^2(\gamma_t)} \left(\sqrt{1-\tanh^2(\gamma_t)} \alpha_t^{y*} - \tanh(\gamma_t) \alpha_t^{x*} \right) \end{array} \right],
\end{aligned}$$

for $t = 1, \ldots, T$. The second derivative is

$$\begin{aligned}
\ddot{p}(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta}) &= \frac{\partial^2 \log p(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})}{\partial \boldsymbol{\theta}_t \partial \boldsymbol{\theta}_t'} \\
&= \frac{\partial \dot{p}(\boldsymbol{z}_t|\boldsymbol{\theta}_t;\boldsymbol{\beta})}{\partial \boldsymbol{\theta}_t'} \\
&= \frac{\partial(\dot{\boldsymbol{Z}}_t' \boldsymbol{H}^{-1} \boldsymbol{z}_t - \dot{\boldsymbol{Z}}_t' \boldsymbol{H}^{-1} Z_t(\boldsymbol{\theta}_t))}{\partial \boldsymbol{\theta}_t'} \\
&= \frac{\partial \dot{\boldsymbol{Z}}_t' \boldsymbol{H}^{-1} \boldsymbol{z}_t}{\partial \boldsymbol{\theta}_t'} - \frac{\partial \dot{\boldsymbol{Z}}_t' \boldsymbol{H}^{-1} Z_t(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t'} \\
&= \frac{\partial \dot{\boldsymbol{Z}}_t' \boldsymbol{H}^{-1} \boldsymbol{z}_t}{\partial \operatorname{vec}(\dot{\boldsymbol{Z}}_t)'} \frac{\partial \operatorname{vec}(\dot{\boldsymbol{Z}}_t)}{\partial \boldsymbol{\theta}_t'} - \dot{\boldsymbol{Z}}_t' \boldsymbol{H}^{-1} \dot{\boldsymbol{Z}}_t \\
&= (\boldsymbol{I}_m \otimes \boldsymbol{z}_t' \boldsymbol{H}^{-1}) \ddot{\boldsymbol{Z}}_t - \dot{\boldsymbol{Z}}_t' \boldsymbol{H}^{-1} \dot{\boldsymbol{Z}}_t,
\end{aligned}$$

where $m = 3$ is the dimension of $\boldsymbol{\theta}_t$ and $n = 2$ is the dimension of $\boldsymbol{z}_t$.

$$\begin{aligned}\ddot{\boldsymbol{Z}}_t &= \frac{\partial \operatorname{vec}(\dot{\boldsymbol{Z}}_t)}{\partial \boldsymbol{\theta}_t'} \\ &= \frac{\partial \operatorname{vec}(\dot{\boldsymbol{Z}}_t)}{\partial \boldsymbol{\alpha}_t^{*\prime}} \\ &= \frac{\partial \operatorname{vec}(\dot{\boldsymbol{Z}}_t)}{\partial \left(\alpha_t^{y*}, \alpha_t^{x*}, \gamma_t\right)} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \sigma_x(1-\tanh^2(\gamma_t)) \\ 0 & 0 & 0 \\ 0 & 0 & -\sigma_x \tanh(\gamma_t)\sqrt{1-\tanh^2(\gamma_t)} \\ 0 & 0 & 0 \\ \sigma_x(1-\tanh^2(\gamma_t)) & -\sigma_x \tanh(\gamma_t)\sqrt{1-\tanh^2(\gamma_t)} & -\sigma_x\sqrt{1-\tanh^2(\gamma_t)}\left(2\tanh(\gamma_t)\sqrt{1-\tanh^2(\gamma_t)}\alpha_t^{y*} + (1-2\tanh^2(\gamma_t))\alpha_t^{x*}\right) \end{bmatrix},\end{aligned}$$

for $t = 1, \ldots, T$.

Therefore,

$$\begin{aligned}\boldsymbol{A}_t &= -\left[\left((\boldsymbol{I}_m \otimes \boldsymbol{z}_t'\boldsymbol{H}^{-1})\ddot{\boldsymbol{Z}}_t - \dot{\boldsymbol{Z}}_t'\boldsymbol{H}^{-1}\dot{\boldsymbol{Z}}_t\right)\Big|_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t}\right]^{-1} \\ \boldsymbol{b}_t &= \hat{\boldsymbol{\theta}}_t + \boldsymbol{A}_t\, \dot{\boldsymbol{Z}}_t'\boldsymbol{H}^{-1}(\boldsymbol{z}_t - Z_t(\boldsymbol{\theta}_t))\Big|_{\boldsymbol{\theta}_t=\hat{\boldsymbol{\theta}}_t},\end{aligned}$$

for $t = 1, \ldots, T$. The problem is that $(\boldsymbol{I}_m \otimes \boldsymbol{z}_t'\boldsymbol{H}^{-1})\ddot{\boldsymbol{Z}}_t$ is random because it depends on γ_t and therefore it happens often in practice that $\boldsymbol{A}_t$ is not positive definite. Other approaches as the modified efficient importance sampling (MEIS) of Koopman et al. (2018) or the numerically accelerated importance sampling (NAIS) of Koopman et al. (2015) might be used instead in order to get an expression for $\boldsymbol{A}_t$ and $\boldsymbol{b}_t$.

4.C.2 Simulation smoothing algorithm

1. Compute $\hat{\boldsymbol{\theta}}_t = \mathrm{E}(\boldsymbol{\theta}_t|\boldsymbol{z}_t;\boldsymbol{\beta})$ by mode estimation, for $t = 1, \ldots, T$.
2. Initialize $\boldsymbol{\theta}_1^+ = \boldsymbol{Z}\boldsymbol{\alpha}_1^+,$ with draw $\boldsymbol{\alpha}_1^+ \sim N(\mathbf{0}, \boldsymbol{P}_1)$.
3. Draw $\boldsymbol{\theta}_t^+$ from $g(\boldsymbol{\theta}_t;\boldsymbol{\beta})$, i.e. the transition equation of the approximate linear Gaussian model (4.4.3), by following the steps:
    a) Draw $\boldsymbol{\eta}_t^+ \sim N(\mathbf{0}, \boldsymbol{Q})$, for $t = 1, \ldots, T$.
    b) Obtain recursively $\boldsymbol{\alpha}_{t+1}^+ = \boldsymbol{T}\boldsymbol{\alpha}_t^+ + \boldsymbol{\eta}_t^+$, and $\boldsymbol{\theta}_t^+ = \boldsymbol{Z}\boldsymbol{\alpha}_t^+$, for $t = 1, \ldots, T$.
4. Use $\boldsymbol{\theta}_t^+$ to generate $\boldsymbol{z}_t^+ \sim g(\boldsymbol{z}_t|\boldsymbol{\theta}_t^+;\boldsymbol{\beta})$, i.e. from the measurement equation of the approximate linear Gaussian model (4.4.3), by following the steps:
    a) Draw $\boldsymbol{\varepsilon}_t^+ \sim N(\mathbf{0}, \boldsymbol{A}_t)$, for $t = 1, \ldots, T$, with $\boldsymbol{A}_t$ evaluated at the mode $\hat{\boldsymbol{\theta}}_t$.
    b) Obtain recursively $\boldsymbol{z}_t^+ = \boldsymbol{\theta}_t^+ + \boldsymbol{\varepsilon}_t^+$, for $t = 1, \ldots, T$.
5. Compute $\hat{\boldsymbol{\theta}}_t^+ = \mathrm{E}(\boldsymbol{\theta}_t|\boldsymbol{z}_t^+;\boldsymbol{\beta})$ by KFS applied to the approximate linear Gaussian model (4.4.3), with $\boldsymbol{A}_t$ evaluated at the mode $\hat{\boldsymbol{\theta}}_t$, and with $\boldsymbol{b}_t = \boldsymbol{z}_t^+$, for $t = 1, \ldots, T$.
6. Compute $\tilde{\boldsymbol{\theta}}_t = \hat{\boldsymbol{\theta}}_t + \boldsymbol{\theta}_t^+ - \hat{\boldsymbol{\theta}}_t^+$, for $t = 1, \ldots, T$; $\tilde{\boldsymbol{\theta}}_t$ is a draw from $g(\boldsymbol{\theta}_t|\boldsymbol{z}_t;\boldsymbol{\beta})$.

5

The spatial lag state space model with an application to regional concentrations of NO_2 in the Netherlands

Abstract

Dutch emissions of nitrogen per hectare are the highest in Europe. Up to 40% of these emissions are made of nitrogen oxides (NO_X), and are due to combustion processes, with road vehicles being their primary contributors. NO_X is an air pollutant that is responsible for the creation of acid rains and of other secondary aerosols and pollutants, and can be harmful for the respiratory system. As such, it is a pollutant of concern and modeling its concentrations in the Netherlands, which are approximated by concentrations of nitrogen dioxide (NO_2), is of interest. This chapter does so at the regional level, by means of a novel spatial lag state space model that takes into account the determinants of NO_2 concentrations, such as traffic intensity and meteorological factors, but also spatial spillovers of NO_2 that are due to its transportation by the wind. The model also has the flexibility of allowing for time-varying coefficients, it is efficiently estimated by Kalman filtering/smoothing and maximum likelihood, and it is employed for forecasting NO_2 concentrations based on different scenarios of road traffic. We find a static effect of the latter variable on the levels of NO_2 concentration, but we also find time-varying differences in NO_2 concentrations between peripheral Dutch regions and inland Dutch regions. Our proposed model realistically forecasts an overall decrease in NO_2 concentrations of around 35%, following an hypothetical 100% decrease in traffic intensity with respect to its actual observed values. The decrease in NO_2 concentrations is gradual and it is predicted to take around eight months for them to achieve their new steady level.

5.1 Introduction

The Netherlands is the region with the largest density of people, animals and economic activity in Europe. The Netherlands Organisation for applied scientific research (TNO, 2019) estimates that Dutch emissions of nitrogen per hectare are the highest in Europe (almost four times the average value), and are made up of 60% ammonia (NH_3) and 40% nitrogen oxides ($NO_X = NO + NO_2$) emissions. NH_3 emissions are mainly caused by agricultural activity, whereas NO_X is emitted during combustion processes. The latter emissions can be from natural sources, such as wildfires and lightning strikes, but for the vast majority they are anthropogenic. The Dutch Government has recently (in 2020) taken actions in both the agriculture and the road transport sectors, in order to reduce nitrogen emissions. Measures to cut NO_X emissions include the reduction of the speed limit on highways, since vehicles are more fuel-efficient at moderate speeds, and the stimulation to use modern and more environmental-friendly vehicles in place of old ones. The focus of this chapter is on NO_X.

Again TNO (2019) estimates that in the Netherlands the largest anthropogenic sources of this air pollutant are represented by road traffic (31%), industry (20%), agriculture (21%) and inland shipping (11%). Road traffic emissions are dominated by diesel consumption of light vehicles (including passenger cars and vans; 14%) and heavy vehicles (trucks and buses; 12%). NO_X emissions in the agricultural sector are due to fertilised soils (15%), natural gas use in horticulture (3.2%) and vehicles (3.2%). Electricity production, together with space heating and wood burning in households and offices, account for the remaining percentage. NO_X emissions are therefore higher in urban areas, and it is estimated that around 90% of Dutch NO_X emissions are exported abroad. All the estimates mentioned above are based on data gathered from the Dutch emission registry, and collected from ground stations in the Netherlands.

At emission NO_X is made up of about 90%-95% nitric oxide (NO), and of nitrogen dioxide (NO_2) for the remaining part. Once in the atmosphere NO is quickly (over a couple of seconds or hours) converted into NO_2. There are also other chemical reactions that take place in the atmosphere. For instance, with the presence of sunlight NO_2 turns back into NO, but it also contributes to the creation of ozone (Finlayson-Pitts & Pitts Jr., 1993). Additionally, NO_X is responsible for the formation of secondary (i.e., not directly emitted in the atmosphere, but created by chemical reactions) aerosols. Stratospheric ozone (i.e., located in the higher level of the atmosphere) and aerosols affect the climate by scattering solar radiation and cooling the Earth's surface (Krotkov et al., 2017). When ozone is instead present in the troposphere (which is the lower level of the atmosphere) it forms smog and hence becomes toxic for human beings (Buchdahl, 2002). Moreover, NO_X contributes to the creation of particulate matter, which is another pollutant, and of acid rains, which contain nitric acid and may be harmful for terrestrial and aquatic ecosystems (Krotkov et al., 2017). Finally, NO_X can have negative effects on human health, since at high concentrations it can cause inflammation of the airways, reduced lung function and increasing susceptibility to respiratory infection (European Environment Agency, 2018). Ogen (2020) also finds a strong relation between long-term exposure to NO_2 and COVID-19 fatalities.

Because of all these side effects, NO_X is an air pollutant of concern and modeling its concentration in the atmosphere over time, by means of econometric techniques, is of interest. Such a model should ideally take into account most of the interactions mentioned above: from the determinants of NO_X emissions, to meteorological conditions that affect its concentration, and spatial spillovers due to its transportation by the wind. In practice this goal is difficult to achieve because of the unavailability of data. For instance, satellite-based data are able to provide rather geographically detailed estimates of pollutants' concentrations, but they are not available for all pollutants, such as NO_X itself. NO_2, on the contrary, is easily measured by remote-sensing techniques and satellite-based NO_2 data have therefore often been used as proxy for NO_X (Duncan et al., 2014). We will also do so in this chapter. However, it should still be kept in mind that the two chemicals are not the same. More NO_X generally implies more NO_2, but if we take as an example the role of sunlight discussed above, we will observe less NO_2 in sunnier periods, which does not imply a lower presence of NO_X and pollution in general; it only means that NO_2 is converted back into NO, which does not diminish the overall concentration of NO_X. It is also challenging finding data that quantify all economic activities responsible for NO_X emissions. In this chapter we employ a dynamic multivariate econometric model for the concentrations of NO_2 in the regions of the Netherlands, which accounts for (time-varying) effects of road vehicles and meteorological conditions on such concentrations, as well as time-varying differences between border and inland regions, spatial dependences and inertial dynamics. The model is also used for forecasting NO_2 concentrations, based on different (hypothetical) scenarios of traffic intensity.

This econometric model is called a spatial lag state space model, and is new to the econometric literature. Spatial econometric models are multivariate regression models that assume a spatial dependence among the cross-sectional/spatial units of a dataset. This spatial dependence can either be incorporated in the error term of the regression, in the regressors, or in the dependent variable. In the latter case it is assumed that the dependent variable of each cross-sectional unit depends on the same variable in neighbouring units. Such a model is known as a spatial lag model, and can be estimated by maximum likelihood.

In the context of time series data, so when the spatial units are repeatedly observed over time for a relatively long time-span, Yu et al. (2008) establish the asymptotic properties of quasi maximum likelihood estimators for large cross-sectional and time dimensions. However, their approach does not take into account an important feature that often characterises time series data: time-varying parameters. These can, for instance, govern the relationships between time series, which would then be captured by time-varying regression coefficients. Time-changing parameters can also be used to let means and variances of series vary over time, and allow us to capture nonlinearities that often affect environmental variables (Nordhaus, 2013; Castle & Hendry, 2020). Blasques et al. (2016) and Catania and Billé (2017) allow for different kinds of time-varying parameters in a spatial lag model, and propose a score-driven approach to estimate them. This method has the drawback of not being very efficient to implement.

State space models are time series models that contain time-varying parameters (which are in this context also called state variables), and that allow them to be estimated in a much more efficient way than the score-drive approach. When the state space model is linear, which occurs when only the regression coefficients or the time series' means are assumed to vary stochastically over time, these can be efficiently estimated by the Kalman filter/smoother. Examples of state space models with a spatial structure only on the state variables or the error term of the model, can be found in Bocci and Petrucci (2016) and Hinrichsen and Holmes (2009). State space and spatial lag econometric models have, however, so far not met in the literature. In the empirical application of this chapter, the regional concentrations of NO_2 in the Netherlands represent the dependent variable of the model. The spatial lag (instead of a spatial error) structure is in this case needed since spatial spillovers of pollutants are due their transportation by the wind, and therefore directly depend on their levels in neighbouring regions. The time-varying effects of traffic intensity discussed above, as well as other state variables, are estimated by the Kalman filter/smoother.

The analysis of NO_2 concentrations and NO_X emissions has received attention from the geoscientific literature. For example, van der A et al. (2006) analyse time series of NO_2 concentrations over China by means of deterministic components. Dallmann and Harley (2010) and Russell et al. (2012) apply non-econometric methods to estimate NO_X emissions from (non) mobile sources in the US.

The effects of economic activity on the chemistry of the atmosphere also represents a subject of interest for both the geoscientific and the economic research communities. Among many others, see Ozturk (2015) who employs panel econometric techniques to measure the impact of energy consumption on emissions of greenhouse gasses for a set of countries, or Bennedsen et al. (2021) who model, forecast and nowcast carbon dioxide emissions in the US using macroeconomic predictors by means of state space models, as well as the entire literature on the environmental Kuznets curve hypothesis, which stipulates an inverted U-shape relationship over time between economic growth and pollution (e.g., Maddison (2006), Wagner (2015) and Lin and Reuvers (2020b)).

Spatial dependencies are of key importance in the analysis of environmental data. In what follows we cite some of the existing studies which take such a data feature into account. Anselin and Le Gallo (2006) conduct a cross-sectional spatial econometric analysis of ozone observations in southern California. Cole et al. (2013) use a similar econometric technique to model carbon dioxide emissions of firms in Japan. Deb and Tsay (2019) develop a spatio-temporal model with space-time interaction for particulate matter in Taiwan. Guan et al. (2019) employ mobile monitors on Google Street View vehicles (in Oakland, California) in a spatio-temporal model, to make short-term forecasts and high-resolution maps of current NO_2 levels.

State space models have already been used, in the econometric literature, for modeling time series of environmental variables. Proietti and Hillebrand (2017) employ them to model seasonal changes in central England temperatures, Bennedsen et al. (2019) to analyse the trend of the airborne fraction and sink rate of anthropogenically released carbon dioxide, Li et al. (2020) to forecast El Niño events, and Hillebrand et al. (2020) to study the relation between global mean sea level and surface temperature.

This chapter lies at the intersection of the four above-mentioned research fields, and it therefore contributes to all of them. It is indeed the first time that a spatial state space model is employed to analyse and forecast NO_2 concentrations, while modeling the (time-varying) effect traffic intensity has on them. Moreover, the novel econometric method here proposed also brings a methodological contribution to the literature on time series and spatial econometrics. Lastly, such regional econometric analyses of NO_2 concentrations have, to the best of our knowledge, not been performed for the Netherlands yet.

The structure of the chapter is as follows. Section 5.2 introduces the spatial lag state space model and explains its estimation procedure. Section 5.3 reports the results of a Monte Carlo simulation study, which is conducted in order to assess the performance of the Kalman smoother and maximum likelihood methods in estimating the proposed econometric model. Section 5.4 discusses the results of an empirical application of the spatial lag state space model for modeling and forecasting Dutch regional concentrations of NO_2.

5.2 The spatial lag state space model

The spatial lag (also known as spatial autoregressive or spatial Durbin) econometric model takes the following form:

$$\boldsymbol{y}_t = \rho \boldsymbol{W} \boldsymbol{y}_t + \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T, \tag{5.2.1}$$

where $\boldsymbol{y}_t$ is the stationary (i.e., mean-reverting) $n \times 1$ vector for the dependent variable, $\boldsymbol{X}_t$ is a known $n \times k$ matrix of stationary regressors (which could also include time lags of $\boldsymbol{y}_t$) with corresponding $k \times 1$ coefficient vector $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}_t$ is the $n \times 1$ vector of errors. The dimensions of the spatial (cross-sectional) and time units are represented by n and T, respectively. The so-called spatial lag of the dependent variable, $\boldsymbol{W}\boldsymbol{y}_t$, is also included in model (5.2.1) as additional regressor, and it is pre-multiplied by the spatial parameter, ρ, which measures the degree of spatial dependence in $\boldsymbol{y}_t$. The spatial weight matrix, $\boldsymbol{W}$, is pre-specified and defines which spatial units are neighbours of each other. The identification of the model is guaranteed by the diagonal elements of $\boldsymbol{W}$, w_{ii}, being equal to zero. Its off-diagonal elements, w_{ij}, are not equal to zero if spatial unit j is a neighbour of spatial unit i, and therefore y_j affects y_i; otherwise $w_{ij} = 0$, for $i, j = 1, \ldots, n$. The spatial weight matrix is therefore sparse. More details about the form of $\boldsymbol{W}$ will be given later in this section.

The regressors $\boldsymbol{X}_t$ are treated as weakly exogenous, but the same cannot be assumed for $\boldsymbol{W}\boldsymbol{y}_t$. Namely, $\mathrm{E}[(\boldsymbol{W}\boldsymbol{y}_t)'\boldsymbol{\varepsilon}_t] \neq 0$ poses an endogeneity problem which renders the ordinary least squares estimators of the parameters of model (5.2.1) inconsistent. Other methods, such as spatial two-stage least squares, or the generalised method of moments estimator, which often make use of the external instruments $\{\boldsymbol{W}\boldsymbol{X}_t, \boldsymbol{W}^2\boldsymbol{X}_t, \ldots\}$ for $\boldsymbol{W}\boldsymbol{y}_t$ (Kelejian & Prucha, 1998), can be employed in order to achieve a consistent estimation of the parameters. Alternatively, by imposing a normality assumption on the error terms, $\boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n)$, with $\boldsymbol{I}_n$ being a $n \times n$ identity matrix, and re-writing the model in the following reduced form (i.e., where the right hand side of the equation contains only exogenous variables)

$$\boldsymbol{y}_t = (\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1} \boldsymbol{X}_t \boldsymbol{\beta} + (\boldsymbol{I}_n - \rho \boldsymbol{W})^{-1} \boldsymbol{\varepsilon}_t, \quad t = 1, \ldots, T, \tag{5.2.2}$$

it is possible to obtain the following log-likelihood expression for $\boldsymbol{y}_t$ (Yu et al., 2008)

$$\begin{aligned}\ell = &-\frac{nT}{2}\log(2\pi) - \frac{nT}{2}\log(\sigma_\varepsilon^2) + T\log\left(\det\left(\boldsymbol{I}_n - \rho\boldsymbol{W}\right)\right)\\ &- \frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^{T}\left(\left(\boldsymbol{I}_n - \rho\boldsymbol{W}\right)\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}\right)'\left(\left(\boldsymbol{I}_n - \rho\boldsymbol{W}\right)\boldsymbol{y}_t - \boldsymbol{X}_t\boldsymbol{\beta}\right),\end{aligned}$$

and estimate the model's parameters (i.e., the β coefficients and σ_ε^2) by maximum likelihood. The method that we propose in this chapter will be based on the maximum likelihood approach.

The spatial weight matrix is generally row-standardised (and it will be in this chapter) such that, for each spatial unit, the spatial lagged variable is a weighted average of the values of $\boldsymbol{y}_t$ observed in neighbouring locations, with weights summing to unity. The row-standardisation also allows to control the space of the spatial parameter, since in such case $|\rho| < 1$ (Lee, 2004). Notice that if $\rho = 1$ and $\boldsymbol{W}$ is row-standardised, $\boldsymbol{I}_n - \rho\boldsymbol{W}$ in equation (5.2.2) is singular and cannot be inverted. Moreover, restricting $|\rho| < 1$ allows to rewrite $(\boldsymbol{I} - \rho\boldsymbol{W})^{-1}$ as the following infinite sum: $\boldsymbol{I} + \rho\boldsymbol{W} + \rho^2\boldsymbol{W}^2 + \rho^3\boldsymbol{W}^3 + \dots$. This implies that the spatial effects are decaying over space, i.e., they become milder as the distance between spatial units increases (to see this notice that $\boldsymbol{W}^2$ represents the matrix of the neighbours' neighbours of the spatial units, and that $\rho^2 < \rho$ when $|\rho| < 1$). A similar decaying effect, over time, is observed in autoregressive (AR) models for time series, where the coefficient of the dependent variable's time lag is also restricted to be lower than unity in absolute value. The spatial weight matrix need not be based (only) on geographical distances, as will also be illustrated in more detail in Section 5.4.1. Depending on the type of cross-sectional units at hand, $\boldsymbol{W}$ could also be built according to economic, financial or cultural proximities.

As already mentioned in the Introduction, spatial relationships can (also) be incorporated in the regressors $\boldsymbol{X}_t$ or in the error term. Although the estimation method proposed in this chapter can be employed also for such model specifications, this possibility is not considered here because it is not supported by the empirical application of Section 5.4. We refer to Anselin (1988) and LeSage and Pace (2009) for an extensive treatment of spatial econometric models.

In a time series setting, so when both n and T are large, which is also the case in this chapter, Yu et al. (2008) provide the asymptotic properties of quasi maximum likelihood estimators of the parameters in model (5.2.1). Blasques et al. (2016) and Catania and Billé (2017) let these parameters vary over time, and estimate them with a score-driven approach. This method assumes that the time-varying parameters depend on their past values as well as past values of the score functions, which measure the sensitivity of the log-likelihood with respect to the time-varying parameters. Although being the only existing method to model time-varying parameters in spatial econometric models, it has the drawback of needing analytical expressions for all score functions, which can vary for different parameters, and it is therefore not computationally efficient to implement. In this chapter we propose an alternative and more efficient way to estimate some of these parameters as time-varying, which makes use of state space models.

The spatial lag state space model takes the form

$$\begin{aligned}
\boldsymbol{y}_t &= \rho \boldsymbol{W} \boldsymbol{y}_t + \boldsymbol{X}_t \boldsymbol{\beta}_t + \boldsymbol{Z} \boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{H}) \\
\boldsymbol{\theta}_t &= (\boldsymbol{\beta}_t', \boldsymbol{\alpha}_t')' \\
\boldsymbol{\theta}_{t+1} &= \boldsymbol{T} \boldsymbol{\theta}_t + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim N(\boldsymbol{0}, \boldsymbol{Q})
\end{aligned} \tag{5.2.3}$$

for $t = 1, \ldots, T$, and it is, as all state space models, composed of an observation equation (the first one) and a transition equation (the last one) which specifies the dynamics of the time-varying parameters, $\boldsymbol{\theta}_t$. In model (5.2.3), $\boldsymbol{\alpha}_t$ is a $p \times 1$ vector of time-varying unobserved/latent components which can, for instance, capture inertial dynamics in the error term of the observation equation, as well as modeling time-varying means or time-varying differences between the intercepts of the dependent variable, $\boldsymbol{y}_t$; $\boldsymbol{Z}$ is a $n \times p$ known selection matrix, which links the dependent variable to the latent components; $\boldsymbol{\theta}_t$ is a $(k + p) \times 1$ vector which collects all the time-varying parameters (also called state variables) of the model, and $\boldsymbol{T}$ is a $(k+p)\times(k+p)$ matrix which defines their dynamic structure; $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are, respectively, $n\times 1$ and $(k+p)\times 1$ error terms, with $\boldsymbol{H}=\sigma_\varepsilon^2\boldsymbol{I}_n$ and $\boldsymbol{Q}$ being their covariance matrices. We also introduce the matrix $\boldsymbol{Z}_t^*=(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\,[\boldsymbol{X}_t \quad \boldsymbol{Z}]$, where $[\boldsymbol{X}_t \quad \boldsymbol{Z}]$ is a $n\times(k+p)$ matrix which concatenates $\boldsymbol{X}_t$ and $\boldsymbol{Z}$. From this point onwards, we will indicate with "hyperparameters" ρ, σ_ε^2, and the parameters of $\boldsymbol{T}$ and $\boldsymbol{Q}$, and with "static parameters" all parameters of the model that are constant over time (i.e., the hyperparameters and the β coefficients that are not time-varying).

The reduced form of model (5.2.3) is

$$\begin{aligned}
\boldsymbol{y}_t &= (\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}(\boldsymbol{X}_t\boldsymbol{\beta}_t+\boldsymbol{Z}\boldsymbol{\alpha}_t)+(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t\sim N(\boldsymbol{0},\boldsymbol{H})\\
\boldsymbol{\theta}_t &= (\boldsymbol{\beta}_t',\boldsymbol{\alpha}_t')'\\
\boldsymbol{\theta}_{t+1} &= \boldsymbol{T}\boldsymbol{\theta}_t+\boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t\sim N(\boldsymbol{0},\boldsymbol{Q})
\end{aligned}$$

for $t=1,\dots,T$, and its hyperparameters can be estimated by maximising the log-likelihood

$$\ell=\sum_{t=1}^{T}\ell_t=\sum_{t=1}^{T}\left(-\frac{n}{2}\log(2\pi)-\frac{1}{2}\log(\det\boldsymbol{F}_t)-\frac{1}{2}\boldsymbol{v}_t'\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\right),$$

where the prediction error, $\boldsymbol{v}_t$, and its covariance matrix, $\boldsymbol{F}_t$, are evaluated by the following Kalman Filter recursions

$$\begin{aligned}
\boldsymbol{v}_t &= \boldsymbol{z}_t-\boldsymbol{Z}_t^*\boldsymbol{a}_t\\
\boldsymbol{F}_t &= \boldsymbol{Z}_t^*\boldsymbol{P}_t\boldsymbol{Z}_t^{*'}+(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1}\boldsymbol{H}(\boldsymbol{I}_n-\rho\boldsymbol{W})^{-1'}\\
\boldsymbol{K}_t &= \boldsymbol{T}\boldsymbol{P}_t\boldsymbol{Z}_t^{*'}\boldsymbol{F}_t^{-1}\\
\boldsymbol{a}_{t|t} &= \boldsymbol{a}_t+\boldsymbol{P}_t\boldsymbol{Z}_t^{*'}\boldsymbol{F}_t^{-1}\boldsymbol{v}_t\\
\boldsymbol{P}_{t|t} &= \boldsymbol{P}_t-\boldsymbol{P}_t\boldsymbol{Z}_t^{*'}\boldsymbol{F}_t^{-1}\boldsymbol{Z}_t^*\boldsymbol{P}_t\\
\boldsymbol{a}_{t+1} &= \boldsymbol{T}\boldsymbol{a}_t+\boldsymbol{K}_t\boldsymbol{v}_t\\
\boldsymbol{P}_{t+1} &= \boldsymbol{T}\boldsymbol{P}_t\,(\boldsymbol{T}-\boldsymbol{K}_t\boldsymbol{Z}_t^*)'+\boldsymbol{Q},
\end{aligned} \tag{5.2.4}$$

for $t = 1, \ldots, T$. In recursions (5.2.4), $\boldsymbol{a}_{t|t}$ is the filtered estimate of the state vector, $\boldsymbol{\theta}_t$, and $\boldsymbol{P}_{t|t}$ is its estimated covariance matrix; $\boldsymbol{a}_{t+1}$ is the one-step-ahead prediction for the state vector, with $\boldsymbol{P}_{t+1}$ being the corresponding predicted covariance matrix. The Kalman filter requires an initialisation for $\boldsymbol{a}_1 = \mathbf{0}$ and $\boldsymbol{P}_1$, which is a diagonal matrix whose diagonal elements are equal to the diffuse or unconditional variances of the elements of $\boldsymbol{\theta}_t$ (more details about the form of $\boldsymbol{P}_1$ are provided in Section 5.3). We refer to Harvey (1989) and Durbin and Koopman (2012) for a detailed discussion of state space models and their estimation.

The estimation accuracy of the state vector can be further improved with the Kalman smoother. The Kalman smoother recursions of De Jong (1989) are

$$\begin{aligned}
\boldsymbol{r}_{t-1} &= \boldsymbol{Z}_t^{*\prime} \boldsymbol{F}_t^{-1} \boldsymbol{v}_t + (\boldsymbol{T} - \boldsymbol{K}_t \boldsymbol{Z}_t^*)' \, \boldsymbol{r}_t \\
\hat{\boldsymbol{\theta}}_t &= \boldsymbol{a}_t + \boldsymbol{P}_t \boldsymbol{r}_{t-1} \\
\boldsymbol{N}_{t-1} &= \boldsymbol{Z}_t^{*\prime} \boldsymbol{F}_t^{-1} \boldsymbol{Z}_t^* + (\boldsymbol{T} - \boldsymbol{K}_t \boldsymbol{Z}_t^*)' \, \boldsymbol{N}_t \, (\boldsymbol{T} - \boldsymbol{K}_t \boldsymbol{Z}_t^*) \\
\boldsymbol{V}_t &= \boldsymbol{P}_t - \boldsymbol{P}_t \boldsymbol{N}_{t-1} \boldsymbol{P}_t,
\end{aligned} \tag{5.2.5}$$

for $t = T, \ldots, 1$, with $\boldsymbol{r}_T = \mathbf{0}$ and $\boldsymbol{N}_T = \mathbf{0}$. The Kalman smoother estimate of $\boldsymbol{\theta}_t$ is represented by $\hat{\boldsymbol{\theta}}_t$, and $\boldsymbol{V}_t$ is its smoothed covariance matrix.

Notice that since the time-varying parameters have their own sources of error in the transition equation, they are assumed to vary stochastically over time. Stochastic specifications are generally very flexible in modeling the evolution of parameters over time. In order to ensure that the stationarity property of $\boldsymbol{y}_t$ is maintained, we are generally imposing an autoregressive dynamic structure for the unobserved components, by assuming that the corresponding diagonal elements in $\boldsymbol{T}$ are below 1 in absolute value. However, there is no intercept in the autoregressive equations, which is equivalent to assuming that the latent variables are returning to zero (which is their implied mean in case of no intercept). This assumption is unlikely to always hold in practice. However, simulation results, that we do not report in this chapter, show that the Kalman filter/smoother estimates of the unobserved components are very robust to such misspecifications. This implies that even when the true intercepts of the autoregressive structures are different from zero, then the autoregressive parameter may not be estimated as being lower than unity, but the resulting Kalman filter/smoother estimates will still be able to follow the true time-variation of the unobserved components. The long-horizon forecasts of the unobserved components, based on the Kalman filter recursions, are instead not robust to such a misspecification, as will be seen in Section 5.4.

For $\boldsymbol{\beta}_t$ we impose a random walk dynamic, which is in line with the econometric literature on stochastically time-varying regression coefficients (see, for instance, Koop and Korobilis (2013)), and ensures the stationarity in $\boldsymbol{y}_t$ unless one of the regressors is constant over time.

Contrary to Blasques et al. (2016) and Catania and Billé (2017), in this chapter we do not allow for stochastic time variation in the spatial parameter nor in the error variances of model (5.2.3). Doing this will make the model become nonlinear and its estimation by Kalman filtering and maximum likelihood infeasible. More sophisticated methods to estimate such models are needed, and their performance is currently under investigation by the author.

5.3 Monte Carlo simulation study

This section reports the results of a Monte Carlo simulation study, which is performed in order to show the performance of the Kalman smoother and maximum likelihood method in estimating, respectively, the state vector (which includes the time-varying coefficients and the latent variables) and the hyperparameters of the spatial lag state space model (5.2.3).

We consider the following five model specifications (data generating processes) for $\boldsymbol{y}_t$, which mainly aim at illustrating some possible advantages that the inclusion of latent variables can have in modeling the dynamics of $\boldsymbol{y}_t$. They are all based on model (5.2.3), with $\rho = 0.7$, $\sigma_{\boldsymbol{\varepsilon}} = 1$, $n = 40$, $T = \{80, 300\}$, $k = 4$ and only the first element of $\boldsymbol{\beta}$ is time-varying: $\boldsymbol{\beta}_t = (\beta_{1,t}, \beta_2, \beta_3, \beta_4)'$, with $\beta_2 = \beta_3 = \beta_4 = 2$. The regressors $\boldsymbol{X}_t \sim N(\mathbf{0}, \boldsymbol{I}_k)$, for $t = 1, \ldots, T$.

1. Time varying coefficients and no unobserved components: $\boldsymbol{\alpha}_t$ and $\boldsymbol{Z}$ are not part of model (5.2.3); $\boldsymbol{T} = \boldsymbol{I}_4$ and $\boldsymbol{Q} = \mathrm{diag}(\sigma^2_\beta, 0, 0, 0)$, with $\sigma_\beta = 1$, and where the diag function creates a diagonal matrix whose diagonal elements are reported in the argument of the function; $\boldsymbol{W}$ is a queen spatial weight matrix with one order of contiguity, which is row-standardised, and whose off-diagonal elements differ from zero if two spatial units (polygons) have a side or a vertex in common[1]; $\boldsymbol{P}_1 = \kappa \boldsymbol{I}_k$.

2. Common time-varying mean: α_t is a scalar ($p = 1$) and it is common among all elements of $\boldsymbol{y}_t$; $\boldsymbol{Z} = \boldsymbol{\imath}_n$ is a $n \times 1$ vector of ones; $\boldsymbol{T} = \mathrm{diag}(1, 1, 1, 1, \gamma_\alpha)$ and $\boldsymbol{Q} = \mathrm{diag}(\sigma^2_\beta, 0, 0, 0, \sigma^2_\alpha)$, with $\gamma_\alpha = 0.5$ and $\sigma_\beta = \sigma_\alpha = 1$; $\boldsymbol{W}$ is a queen spatial weight matrix with one order of contiguity; $\boldsymbol{P}_1 = \mathrm{blockdiag}\left(\kappa \boldsymbol{I}_k, \frac{\sigma^2_\alpha}{(1-\gamma^2_\alpha)}\right)$, where the blockdiag function creates a block-diagonal matrix whose diagonal blocks are reported in the argument of the function.

3. Time-varying dummies: $\boldsymbol{\alpha}_t$ is a 2×1 vector ($p = 2$) and captures time-varying differences between the intercepts of the dependent variable, $\boldsymbol{y}_t$; $\boldsymbol{Z} = [\mathrm{blockdiag}(\boldsymbol{\imath}_{10}, \boldsymbol{\imath}_{10})' \;\; \boldsymbol{0}_{2\times 20}]'$, with $\boldsymbol{0}_{2\times 20}$ being a 2×20 matrix of zeros; $\boldsymbol{T} = \mathrm{diag}(1, 1, 1, 1, \gamma_{\alpha,1}, \gamma_{\alpha,2})$ and $\boldsymbol{Q} = \mathrm{diag}(\sigma^2_\beta, 0, 0, 0, \sigma^2_{\alpha,1}, \sigma^2_{\alpha,2})$, with $\gamma_{\alpha,1} = \gamma_{\alpha,2} = 0.5$ and $\sigma_\beta = \sigma_{\alpha,1} = \sigma_{\alpha,2} = 1$; $\boldsymbol{W}$ is a queen spatial weight matrix with one order of contiguity; $\boldsymbol{P}_1 = \mathrm{blockdiag}\left(\kappa \boldsymbol{I}_k, \mathrm{diag}\left(\frac{\sigma^2_{\alpha,1}}{(1-\gamma^2_{\alpha,1})}, \frac{\sigma^2_{\alpha,2}}{(1-\gamma^2_{\alpha,2})}\right)\right)$.

4. Deterministic time-varying weight matrix: the only difference from specification 3 above is that the structure of $\boldsymbol{W}_t$, although still being pre-specified, changes at each point in time. The time-variation in $\boldsymbol{W}_t$ is therefore deterministic (non-stochastic), contrary to the other time-varying parameters of the model. The values for $\boldsymbol{W}_t$, for $t = 1, \ldots, 80$, are obtained based on the contiguities of the 40 Dutch COROP regions and data about wind speed and direction, as described in Section 5.4.1. In this case we do not also consider $T = 300$ since there is no available data for such a large sample size.

[1] Such a matrix is built based on the contiguities of the 40 Dutch COROP regions, which are the spatial units employed in the empirical application of Section 5.4.

5. Moving averages: $\boldsymbol{\alpha}_t$ is a $n \times 1$ vector $(p = n)$, such that each element of $\boldsymbol{y}_t$ has its own latent variable, and it captures inertial dynamics in the error terms of the observation equation; $\boldsymbol{Z} = \boldsymbol{I}_n$, $\boldsymbol{T} = \text{blockdiag}\left(\text{diag}(1,1,1,1), \gamma_\alpha \boldsymbol{I}_n\right)$ and $\boldsymbol{Q} = \text{blockdiag}\left(\text{diag}(\sigma_\beta^2, 0, 0, 0), \sigma_\alpha^2 \boldsymbol{I}_n\right)$, with $\gamma_\alpha = 0.5$ and $\sigma_\beta = \sigma_\alpha = 1$; $\boldsymbol{W}$ is a queen spatial weight matrix with one order of contiguity; $\boldsymbol{P}_1 = \text{blockdiag}\left(\kappa \boldsymbol{I}_k, \frac{\sigma_\alpha^2}{(1-\gamma_\alpha^2)} \boldsymbol{I}_n\right)$.

Notice that diagonal elements equal to 1 in $\boldsymbol{T}$, coupled with zero diagonal elements in $\boldsymbol{Q}$, imply a time-constant specification and Kalman smoother estimation of the corresponding state variable. On the other hand, all the time-varying parameters have an autoregressive coefficient different from zero (equal to 1 for $\beta_{1,t}$ and to 0.5 for the unobserved components) and an error variance equal to unity. The value for κ is chosen to be large (it is equal to 100 in the Monte Carlo simulation study and to 10 in the empirical application of Section 5.4, as these values ensure the convergence of the Kalman filter/smoother due to the small magnitude of the β coefficients), since a diffuse initialisation of the Kalman filter is needed for the non-stationary time-varying parameter $\beta_{1,t}$.

The maximisation of the log-likelihood requires an initialisation for the hyperparameters of the model. All hyperparameters are initialised at their true values, in order to reduce the computational time needed to obtain the results. However, simulation results, that we do not include in this chapter, show that the maximum likelihood estimates are not very sensitive to the initial values.

The results that are about to be discussed are based on 500 Monte Carlo simulations. Figures 5.A.1-5.A.4 and 5.A.6 display the distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, of model specifications 1-5, respectively, and when $T = 80$. All estimators seem to be unbiased since their distributions are centered around the corresponding true values of the parameters. Figures 5.A.7-5.A.10 report the same results, for model specifications 1-3 and 5, respectively, obtained with a larger sample size of $T = 300$. As expected, the estimation accuracy of the estimators improves as all distributions become more bell-shaped and concentrated around the true values of the static parameters. Finally, Figure 5.A.5 shows the accurate performance of the Kalman smoother in estimating the time-varying parameters of model specification 4, when $T = 80$.

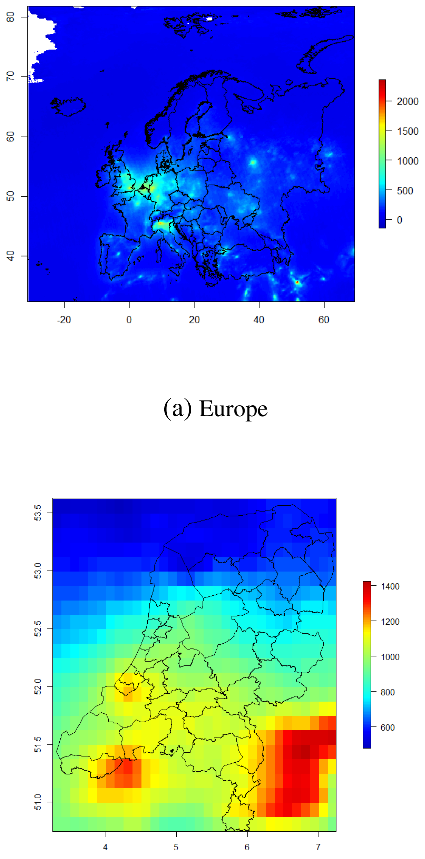
5.4 Empirical application to regional concentrations of NO_2 in the Netherlands

5.4.1 Data description

The data employed in the empirical analysis are about tropospheric NO_2 concentrations, road vehicles as a pollutant emission source, and meteorological conditions. Data about traffic intensity are available monthly from January 2011 until December 2017 ($T = 84$), and for the ($n = 40$) Dutch regions defined by the "Coördinatiecommissie Regionaal Onderzoeksprogramma" (COROP), which represent the spatial units of the analysis. All the other variables employed in the econometric model are therefore, if necessary, spatially aggregated to the COROP regional level, temporally aggregated to the monthly frequency and limited to the aforementioned time span. Figure 5.B.1 displays the maps of the Dutch provinces and COROP regions (the latter are part of the former).

Data for tropospheric NO_2 concentrations

There are two main possible data sources for pollutants' concentrations: measurement stations and satellite pictures. Although the former provide more precise estimates, they are spatially sparse. The latter are less accurate but provide a much higher spatial coverage (Fowlie et al., 2019). Since spatial variation is of key importance in our study, we employ satellite-based data for NO_2 concentrations. TNO (2019) also pushes towards the use of satellite-based data in order to obtain more timely and geographically detailed estimations of pollutants' concentrations, and to measure import and export of emissions (although they do not consider satellite-based data as a substitute for ground measurements).



(a) Europe

(b) The Netherlands

Figure 5.4.1: Tropospheric NO_2 concentrations in Europe and the Netherlands, averaged over the entire monthly series, measured by OMI and based on the DOMINO v2.0 retrieval. Unit of measure: 10^{13}molec./cm^2.

A detailed description of available satellite-based data for atmospheric pollution can be found in Duncan et al. (2014). In what follows we summarise the type of satellite-based data that we employ in the empirical analysis, based on the aforementioned paper. Satellite instruments are able to detect the number of molecules of a particular gas between the instrument and the Earth's surface, which is technically called "vertical column density" (VCD), measured in units of molecules per unit area of the Earth's surface. The Ozone Monitoring Instrument (OMI) resides on the polar-orbiting Aura satellite, which was launched in July 2004. Since the 1st of October of the same year, the instrument has been collecting daily data at approximately the same local time (early afternoon) at every location of the globe. Such passive instruments detect the solar radiation that is backscattered by the Earth's atmosphere and surface. This radiation passes through spectrometers, which are devices that measure energy intensity as a function of wavelength, and allow creation of a spectrum of wavelengths. Each pollutant has a unique way to absorb and reflect specific wavelengths. By comparing the spectral signature recorded by the satellite instrument to a spectral laboratory-measured signature for a given quantity of the pollutant, it is then possible to infer the quantity of that pollutant in the atmosphere. This last step is performed with so-called "retrieval algorithms". For NO_2 concentrations, we employ data obtained with the retrieval algorithm based on the second version of the DOMINO product (Boersma et al., 2011), provided by the Tropospheric Emissions Monitoring Internet Service (TEMIS)[2]. The product provides Level 3 data, where the level corresponds to the degree of data processing. Contrary to Level 2, Level 3 data is mapped to a regular spatial grid and averaged over time. It generally has lower spatial resolution than Level 2, but it has the advantage of being easier to read, visualise and analyse. We directly employ the monthly data provided by the DOMINO product. Most of the NO_2 VCD is found near its surface emission sources because its chemical life is short, i.e., hours to about a day

[2] Satellite-based data can be downloaded from http://temis.nl. An alternative retrieval algorithm is the third version of the NASA standard product (Krotkov et al., 2017), performed by the National Aeronautics and Space Administration Goddard Space and Flight Center (NASA-GSFC) and available on https://disc.gsfc.nasa.gov. The two algorithms produce very similar atmospheric quantities. However, the DOMINO product offers a better resolution in terms of pixel size.

depending on meteorological conditions (Duncan et al., 2014). The contribution to the NO_2 VCD from the stratosphere is already subtracted in Levels 2 and 3 of the OMI NO_2 products, yielding tropospheric NO_2 VCD.

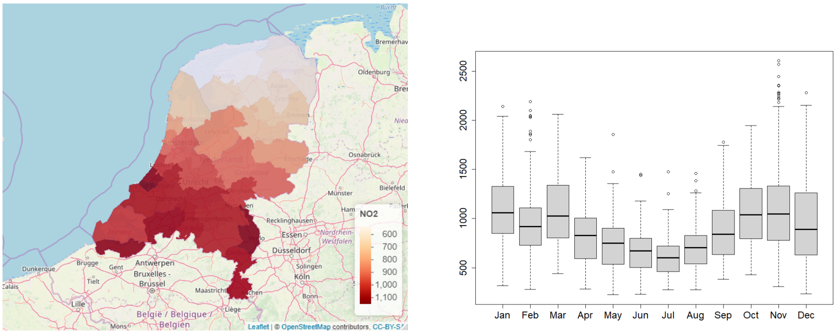
Figure 5.4.1 displays the concentrations of tropospheric NO_2 in Europe and in the Netherlands based on the data described above, and averaged over time. Each pixel, which covers 13×14 km^2 of the Earth's surface, represents a certain value of NO_2 concentration, which is measured in 10^{13}molec./cm^2. The Figure shows that this type of pollution is highest in the Benelux area (which includes Belgium, the Netherlands and Luxembourg) and in Northern Italy, followed by Western Germany and the region around London (UK). Zooming in the Netherlands reveals that the more southern regions of the country and the area around Rotterdam are the most polluted. However, in the former regions, NO_2 seems to have been, at least partially, transported by the wind from the German industrial area delimited by Cologne, Düsseldorf and Dortmund, and the Belgian city of Antwerp. Figure 5.4.2a shows the same data for the COROP regions, which are obtained by averaging the values of the corresponding overlapping pixels. Figure 5.4.2b uses box plots to summarise the distribution of NO_2 concentrations in each month of the year. There is clearly less concentration of NO_2 during warmer seasons, not only because of a lower use of household heating, but also because of the role that some meteorological factors play, such as the amount of sunlight. Such seasonality, together with the international spatial spillovers mentioned above, is taken into account while building the econometric model for NO_2 concentrations.

Data for traffic intensity

Since road vehicles are the primary source of NO_X emissions, data about traffic intensity can potentially be very powerful in modeling NO_2 concentrations. This type of data is collected and provided by Statistics Netherlands[3].

Traffic intensity is measured by counting the number of motor vehicles passed (in both directions) per hour at a (fixed) measuring point on national roads.

[3] Traffic intensity data can be downloaded from the open access database of Statistics Netherlands: https://opendata.cbs.nl.



(a) NO_2 concentrations (unit of measure: 10^{13}molec./cm^2)

(b) Seasonality of NO_2 concentrations

Figure 5.4.2: The left panel displays the satellite-based data about NO_2 concentrations transformed into COROP regional data, averaged over the months. The right panel summarises the distribution of such NO_2 concentrations, in each month.

These counts are performed electronically 24 hours a day every day of the week, by around 20,000 road sensors located on the highways of the Netherlands, whose location is shown in Figure 5.B.5. By multiplying these counts by 24 (number of hours per day) and the number of days in each month, it is possible to obtain the number of road vehicles per month. These values are subsequently divided by the number of hectares that form each COROP region, in order to allow for a fair comparison among spatial units of different geographical dimensions. We indicate with $Traffic$ the variable representing the number of motor vehicles per hectare and per month, passed in each COROP region.

Traffic intensity data are not available for the COROP regions of Oost-Groningen and Noord-Friesland. Missing are also the first three monthly observations for Noord-Overijssel, and the monthly observations for years 2011 and 2015 for Zeeuwsch-Vlaanderen. We impute these missing values with the observed values for the corresponding province (which contains the COROP region at hand, and whose traffic data is obtained as a weighted average of the traffic data of its COROP regions): respectively, Groningen, Friesland, Overijssel, and Zeeland.

Other economic variables, such as energy consumption/production as well as the total number and sizes of farms and industries, would capture the share of NO_X emissions that are due to these other types of economic activity. However, data about these variables are unfortunately publicly available only either at the yearly frequency or at the national level. Because of their heavy lack of spatial or time variation, we do not use these economic indicators in the empirical analysis. Moreover, Curier et al. (2014) employ a chemistry transport model to show that in Benelux, at OMI overpass time, NO_X emissions are mainly caused by combustion processes in the road transport sector and explain almost half of the OMI NO_2 VCD. This finding reinforces the leading role that data about traffic intensity plays in modeling the pollutant. Finally, Lin and Reuvers (2020a) urge econometricians to start modeling air pollutants with variables that are more direct sources of pollution, and not only proxies for economic activity (such as population or gross domestic product). We will therefore not consider such variables as possible regressors in the model for NO_2 concentrations.

Data for the meteorological variables

We consider five meteorological variables, whose data are collected by measurement stations and can be found in the database of the Royal Netherlands Meteorological Institute (in Dutch: Koninklijk Nederlands Meteorologisch Instituut or KNMI)[4] . Namely, temperature ($Temp$), sunlight duration (Sun), amount of rainfall ($Rain$), wind speed ($WindSpd$) and wind direction ($WindDir$). In the Introduction we mentioned the role of sunlight in determining the level of NO_2 concentrations. We do not have any prior knowledge about the effect of temperature and rain, but we can discover it by including these variables in the analysis. The role of these first three meteorological factors is to capture the seasonality of NO_2 concentrations. Data about wind speed and direction are used to determine the spatial dependences among COROP regions, as will be explained in Section 5.4.1.

[4] Meteorological data can be downloaded from http://projects.knmi.nl.

Figures 5.B.2 and 5.B.3 illustrate the transformation process applied to data from 50 measurement stations located throughout the Netherlands, into data for the COROP regions. This step is necessary because the COROP regions represent the spatial units needed for the econometric analysis, and they are spatial polygons. The locations of the measurement stations are instead spatial points. This data transformation is done as follows. First, we check which stations have measured a particular variable. Second, since all meteorological variables are observed at the daily frequency, we aggregate them to the monthly frequency by averaging the daily observations for $Temp$, $WindSpd$ and $WindDir$ (since they are stock variables), and summing the daily observations of Sun and $Rain$ (since they are flow variables), in order to keep their "per month" interpretation. Third, we proceed with the spatial interpolation in order to estimate the values of the variables for the COROP regions. To do so we create a fine spatial square grid that spans the territory of the Netherlands. Then we employ inverse distance weighting in order to predict the variables' values for every cell of the grid. Specifically, let x_i generally define the value of a meteorological variable for measurement station i. Then the predicted/interpolated value of the same variable for grid cell j is obtained as (Anselin & Le Gallo, 2006):

$$x_j = \frac{\sum_i w_i x_i}{\sum_i w_i},$$

with $w_i = 1/(d_{ij}^2)$, where d_{ij} is the distance between station i and grid cell j. The spatial predictions are therefore obtained as weighted averages of the variables' values at all stations, with larger weights given to closer stations. The left panels of Figures 5.B.2 and 5.B.3 display the results from this spatial interpolation. In order to get the final values of the meteorological variables for the COROP regions (right panels of Figures 5.B.2 and 5.B.3), we average the interpolated values of the COROP regions' overlapping grid cells.

A few words need to be spent on the temporal and spatial aggregation of $WindDir$, which represents the direction from which the wind originates. The values for this variable are reported in degrees: 0 or 360 indicate the

North, 90 the East, 180 the South and 270 the West. This implies that averaging these values by taking their sample mean will not yield meaningful results. We therefore aggregate $WindDir$ (either spatially or temporally) by computing its average as follows (Grange, 2014). Let x_i denote the wind direction for observation i. Then the average of $WindDir$ is

$$\bar{x} = \begin{cases} 360 + \left(\arctan\left(\frac{\bar{u}}{\bar{v}}\right) 180\right)/\pi & \text{if } \left(\arctan\left(\frac{\bar{u}}{\bar{v}}\right) 180\right)/\pi \leq 0 \\ \left(\arctan\left(\frac{\bar{u}}{\bar{v}}\right) 180\right)/\pi & \text{otherwise,} \end{cases}$$

where $\bar{u} = \frac{1}{N}\sum_{i=1}^{N} \sin\left(\frac{2\pi x_i}{360}\right)$, $\bar{v} = \frac{1}{N}\sum_{i=1}^{N} \cos\left(\frac{2\pi x_i}{360}\right)$, and N is the sample size at hand. When performing the spatial interpolation discussed above, the latter formulae become $\bar{u} = \frac{\sum_{i=1}^{N} w_i \sin\left(\frac{2\pi x_i}{360}\right)}{\sum_{i=1}^{N} w_i}$ and $\bar{v} = \frac{\sum_{i=1}^{N} w_i \cos\left(\frac{2\pi x_i}{360}\right)}{\sum_{i=1}^{N} w_i}$, and the average is the prediction for grid cell j ($\bar{x} = x_j$).

Based on figures 5.B.2 and 5.B.3, we can conclude that, on average, temperature is, as expected, higher in the South. The wind speed is stronger and the sunlight duration is longer along the coast. The wind is mainly blowing from the South-West, and rainfall is more abundant in the central-western regions of the country. Further details on these meteorological variables, such as measurement units and descriptive statistics, can be found in Table 5.B.1[5].

Time-varying spatial weight matrix based on wind speed and direction

Since spatial spillovers of NO_2 are due to its transportation by the wind, the spatial weight matrix should ideally capture this role of the wind, and therefore be built on information about it. We do so by slightly modifying the approach of Merk and Otto (2020) in using data about wind speed and direction in order to construct a deterministic time-varying spatial weight matrix. Namely, we let

$$\boldsymbol{W}_t = \phi \boldsymbol{W}^{(D)} + (1-\phi)\boldsymbol{W}_t^{(W)} \tag{5.4.1}$$

[5] The standard deviation for $WindDir$ is there calculated with the method of Yamartino (1984), which also employs trigonometric functions.

for $t = 1, \ldots, T$, which implies that $\boldsymbol{W}_t$ is a linear combination of a static and a dynamic spatial weight matrix. The former, $\boldsymbol{W}^{(D)}$, is row-standardised and it is constructed based on geographical distances between spatial units; its diagonal elements are equal to zero and its off-diagonal elements are different from zero if the centroids of the corresponding COROP regions are closer than a certain threshold, which is chosen to make sure that all regions have at least one neighbour. The latter spatial weight matrix, $\boldsymbol{W}_t^{(W)}$, is built on wind characteristics. Let $w_{ij,t}^{(W)}$ be the element corresponding to its i^{th} row and j^{th} column, for $i, j = 1, \ldots, n$. Then the diagonal elements of $\boldsymbol{W}_t^{(W)}$ are equal to zero, whereas its off-diagonal elements are equal to

$$\begin{aligned} w_{ij,t}^{(W)} &= \frac{WindSpd_{j,t}}{d_{ij}} K(\delta_{ij,t}) \\ \delta_{ij,t} &= \frac{2}{r} \left[\mod \left(\frac{b_{ij} - WindDir_{j,t} + 180}{360} \right) - 180 \right], \end{aligned} \tag{5.4.2}$$

for $i, j = 1, \ldots, n$ and $t = 1, \ldots, T$. In the formula above, b_{ij} represents the bearing (i.e., the relative position) of the centroid of COROP region j with respect to the centroid of COROP region i, and d_{ij} is the great circle distance between them (measured in km). The bearing is measured in degrees and varies within the interval $(0, 360]$. For instance, a bearing of 90 degrees indicates that the centroid of COROP region j is located to the east of the centroid of COROP region i. In equation (5.4.2), r represents the range of maximum admissible deviations between the prevailing wind direction in COROP region j at time t, $WindDir_{j,t}$, and the bearing. The more the bearing is aligned with the wind direction, the more $\delta_{ij,t}$, in equation (5.4.2), takes values close to zero. The latter is the argument of a kernel function, K, whose values are positive and decrease while $\delta_{ij,t}$ moves away from zero. We employ the Epanechnikov kernel function, which takes the form

$$K(\delta_{ij,t}) = \begin{cases} \frac{3}{4}\left(1 - \delta_{ij,t}^2\right) & \text{if } \delta_{ij,t} \in [-1, 1] \\ 0 & \text{otherwise,} \end{cases}$$

for $i, j = 1, \ldots, n$ and $t = 1, \ldots, T$. This implies that if the deviation between the bearing and the wind direction exceeds $r/2$, then $w_{ij,t} = 0$ and therefore the concentrations of NO_2 in COROP region i are not transported by the wind from COROP region j, at time t. Moreover, the sparsity of the spatial weight matrix is so guaranteed. On the other hand, for an increasing alignment between the bearing and the wind direction, the kernel function takes larger values and therefore intensifies the spatial dependence of COROP region i on COROP region j. These values are finally multiplied by the average wind speed in COROP region j at time t, $WindSpd_{j,t}$, and divided by d_{ij}, since these two measures respectively strengthen and weaken the spatial relationships.

We refer to Merk and Otto (2020) for a visualisation of the above-described method to build a time-varying spatial weight matrix based on data about wind speed and direction.

In practice we choose $\phi = 0.85$ and $r = 150$, which are the values that Merk and Otto (2020) observe in their empirical study, even though they analyse a different pollutant (particulate matter) in a different geographical location (eastern United States)[6]. However, this choice for r does not prevent some rows of $\boldsymbol{W}_t^{(W)}$ from having only zero elements; we therefore let those rows of $\boldsymbol{W}_t$ be equal just to the corresponding rows of $\boldsymbol{W}^{(D)}$. The non-zero rows of $\boldsymbol{W}_t^{(W)}$, which is built as described above, are then standardised at each point in time, before $\boldsymbol{W}_t^{(W)}$ is plugged in formula (5.4.1). This makes sure that also $\boldsymbol{W}_t$ is row-standardised. On average, around 50% of the elements of $\boldsymbol{W}_t$ are equal to zero.

5.4.2 Empirical results

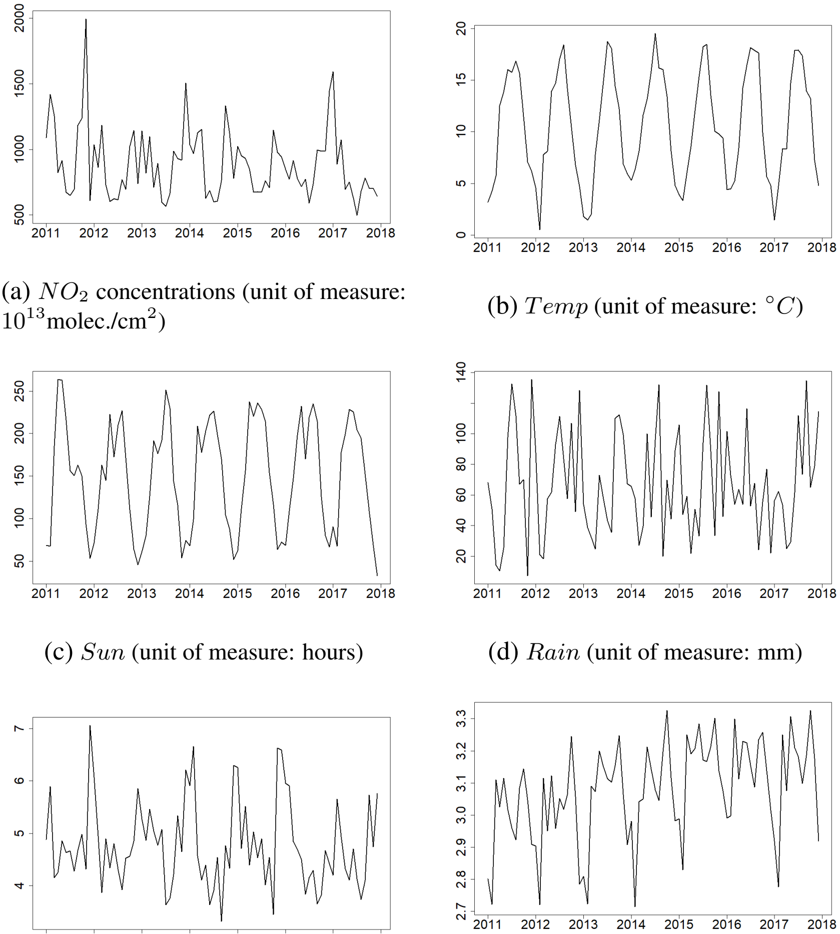
Before diving into the model specification and discussion of the empirical results, we check whether our data satisfy some of the assumptions that we

[6] Merk and Otto (2020) estimate ϕ by standard maximum likelihood and choose, from a grid of values for r, the one that maximises the log-likelihood function, in order to estimate r. Nonetheless, the theoretical properties of these "estimators" have not been derived and we therefore prefer to fix values for the two parameters a priori.

made in Section 5.2. Figure 5.4.3 displays the time series of the variables employed in the empirical analysis, for the Netherlands (i.e., the averaged observed data over the Dutch COROP regions). We notice that all time series, except for $Traffic$, seem to be stationary around a mean. Traffic intensity looks instead stationary around a positive trend[7]. Therefore, despite an increase of road vehicles per hectare over time, we do not notice a similar increase in NO_2 concentrations. This finding can have two main possible explanations: either the increasing efficiency of vehicles and environmental awareness tend to decrease the contribution of traffic to NO_2 concentrations over time, or other types of economic activities are subject to such an environmentally sustainable change, or both. Therefore, modeling the effect of $Traffic$ on NO_2 as time-varying, is of interest in order to shed light on this issue. Figure 5.4.4 displays the time series for the same variables, but this time observed for the six COROP regions in the province of Zuid-Holland. The comovements of these series among neighbouring regions suggests that there is spatial dependence among them, which motivates the use of a spatial lag-type of econometric model.

The model specification for the analysis of regional NO_2 concentrations in the Netherlands, is based on specification 4 of the Monte Carlo simulation study of Section 5.3. Namely, model (5.2.3), where $T = 84$ and $\boldsymbol{y}_t$ is a 40×1 vector ($n = 40$) whose elements correspond to the natural logarithm of NO_2 concentrations in the Dutch COROP regions: $y_{i,t} = \log(NO_{2,i,t})$ for $i = 1, \ldots, 40$. The use of the logarithms instead of the levels for the dependent variable avoids problems such as obtaining negative predicted values for $\boldsymbol{y}_t$, since NO_2 concentrations cannot be negative (this choice is common in the climate econometrics literature: see, e.g., Wagner (2015)). The spatial weight matrix, $\boldsymbol{W}_t$, is time-varying and built as explained in Section 5.4.1. The structure of (partial) autocorrelation functions for the elements of $\boldsymbol{y}_t$ (which we do not report in this chapter) suggests that the dependent vari-

[7] Such stationary behaviour is widely confirmed by the Augmented Dickey-Fuller test (Elliott et al., 1996) for the presence of a unit root (which accounts for a deterministic time trend when appropriate), applied to all regional time series; at the 5% confidence level, only $Traffic$ for the COROP region “Zeeuwsch-Vlaanderen” fails to reject the null hypothesis. We do not report these results in the chapter.



(a) NO_2 concentrations (unit of measure: 10^{13}molec./cm^2)

(b) $Temp$ (unit of measure: $^\circ C$)

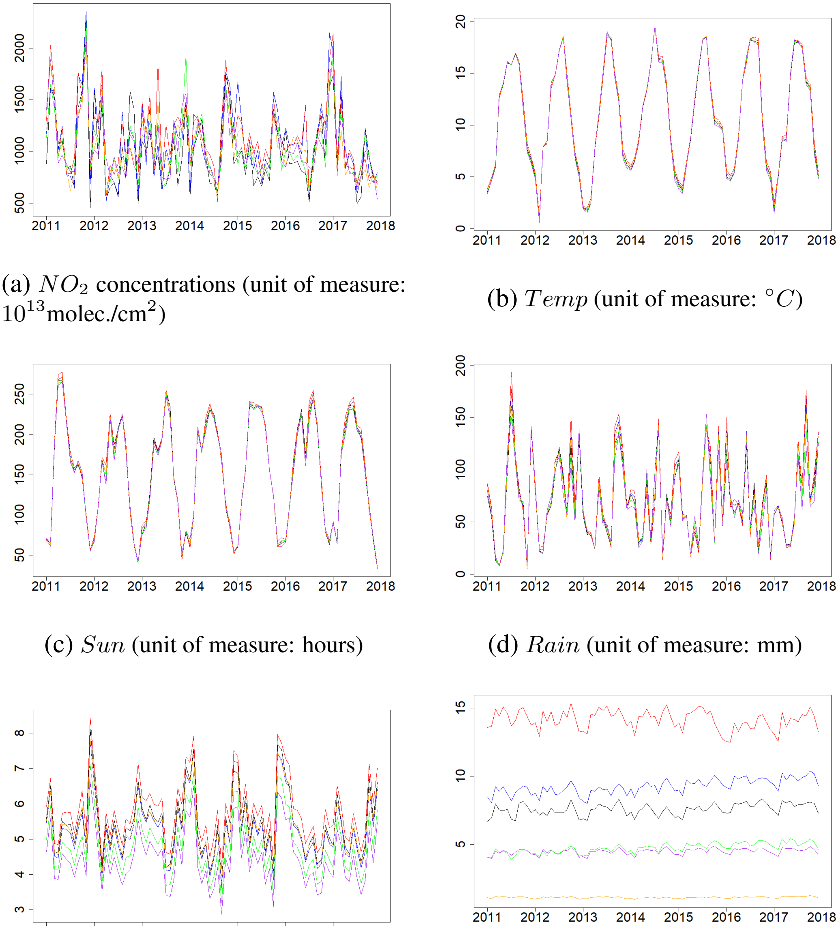
(c) Sun (unit of measure: hours)

(d) $Rain$ (unit of measure: mm)

(e) $WindSpd$ (unit of measure: m/s)

(f) $Traffic$ (unit of measure: number of vehicles/hectare)

Figure 5.4.3: Time series plots of the variables observed for the Netherlands. The circular characteristic of $WindDir$ hampers the visualisation of its time series, which is therefore not here included.



(a) NO_2 concentrations (unit of measure: 10^{13}molec./cm^2)

(b) $Temp$ (unit of measure: $^\circ C$)

(c) Sun (unit of measure: hours)

(d) $Rain$ (unit of measure: mm)

(e) $WindSpd$ (unit of measure: m/s)

(f) $Traffic$ (unit of measure: number of vehicles/hectare)

Figure 5.4.4: Time series plots of the variables observed for the six COROP regions in the province of Zuid-Holland. The circular characteristic of $WindDir$ hampers the visualisation of its time series, which are therefore not here included.

able follows an order one autoregressive rather than a moving average structure. Therefore, the 40×6 ($k = 6$) matrix of regressors, $\boldsymbol{X}_t$, contains an intercept and the following explanatory variables: $\log(NO_{2,t-1})$, $Traffic$, $Temp$, Sun and $Rain$.

The vector of unobserved components, $\boldsymbol{\alpha}_t = (\alpha_{Sea,t}, \alpha_{Border,t})'$, is made of two ($p = 2$) elements which aim to capture time-varying deviations between the intercepts of COROP regions that, respectively, face the sea and are located at land borders, and the ones that do not. This implies that $\alpha_{Sea,t}$ is common among maritime regions, whereas $\alpha_{Border,t}$ is common among regions bordering Belgium and/or Germany, and the 40×2 selection matrix $\boldsymbol{Z}$ is built accordingly. The role of such a latent vector is therefore to account for the so-called "border effect": on one hand the fact the pollution of regions located at land borders can affect and be affected by neighbouring countries; on the other hand the fact that the little pollution coming from the sea (mainly shipping pollution) can only very limitedly influence the NO_2 concentrations of maritime regions, while the pollution of such regions can still be exported to the sea. Another way to account for the border effect would be to include the neighbouring German/Belgian regions and fictitious geographical regions located on the sea, as spatial units in the model, and augment $\boldsymbol{y}_t$ with their levels of NO_2 concentrations. This would imply an extension of the spatial weight matrix, but not necessarily an inclusion of explanatory variables also for these additional regions. However, the use of latent components in order to model the border effect is novel and illustrates an advantage of including a state space structure in the spatial lag econometric model, over employing a more static specification.

We let $\boldsymbol{\beta}_t = (\beta_{Intercept}, \beta_{NO_{2,t-1}}, \beta_{Traffic,t}, \beta_{Temp}, \beta_{Sun}, \beta_{Rain})'$ be the 6×1 coefficient vector, which assumes that only the coefficient of $Traffic$ varies over time. Indeed, although the model could potentially allow for more time-changing coefficients, it makes little intuitive sense to assume that meteorological factors can have a varying effect on NO_2 concentrations over time. The time-varying effect of traffic intensity on pollution was instead motivated at the beginning of this section. Therefore, the transition matrix is $\boldsymbol{T} = \text{blockdiag}(\boldsymbol{I}_6, \text{diag}(\gamma_{\alpha,Sea}, \gamma_{\alpha,Border}))$ and the covariance matrix of the state innovations is $\boldsymbol{Q} = \text{diag}(0, 0, \sigma^2_{Traffic}, 0, 0, 0, \sigma^2_{\alpha,Sea}, \sigma^2_{\alpha,Border})$.

We compare the results that model (5.2.3) yields, to the ones obtained with the spatial lag model (5.2.1), which does not include as many dynamics as the spatial state space model does, and is therefore of a more static nature. Dummy variables for maritime and bordering regions are the static equivalent of the two unobserved components discussed above. Therefore, in this empirical analysis, the matrix of regressors of model (5.2.1) includes the same regressors as model (5.2.3) plus these two dummies (i.e., it is equal to $[\boldsymbol{X}_t \quad \boldsymbol{Z}]$), whose coefficients will be denoted as β_{Sea} and β_{Border}, respectively. We employ the same spatial weight matrix in the two models. The initial values (needed for the maximisation of the log-likelihood) for the parameters that need to be estimated by maximum likelihood, are equal to 1 for all variance parameters, and 0 for the remaining parameters.

The in-sample results based on the entire sample (i.e., January 2011 - December 2017) for both models are reported in Table 5.4.1. All common coefficients between the models are estimated similarly. The standard errors (reported in parentheses in Table 5.4.1) for the Kalman smoother estimates of the static β coefficients of the state space model, correspond to the square roots of the respective diagonal elements of $\boldsymbol{P}_{T|T}$ from recursions (5.2.4)[8]. However, there are only theoretical grounds to test hypotheses on the parameters that are estimated by maximum likelihood.

We find that the spatial dependence, represented by the estimate of ρ, is positive, significant and very large in magnitude, thus suggesting that the inclusion of the spatial structure is appropriate. As expected, the effect of traffic intensity is positive and significant ("ceteris paribus" is here and in what follows left implicit), whereas sunlight yields a negative and significant contribution to NO_2 concentrations. We also find the time lag of NO_2 concentrations and temperature to have a positive and significant impact on the dependent variable, and the coefficient of $Rain$ is estimated as being negative and significant. From the spatial lag model (5.2.1) we can conclude that maritime and border regions significantly have, respectively, lower and higher concentrations of NO_2, with respect to the remaining inland regions. This result is

[8] Since the diagonal elements of $\boldsymbol{P}_{T|T}$ that refer to constant state variables, are equal to the diagonal elements of $\boldsymbol{V}_t$, from recursions (5.2.5), for $t = 1, \ldots, T$.

also to be expected: as Figure 5.4.1b shows, most of the border regions are located in proximity of industrial and very polluted Belgian and German locations, therefore importing some of this pollution and increasing their overall level of NO_2 concentrations; on the contrary, the mostly clean air coming from the sea has a reducing effect on the quantity of pollution in maritime regions.

When the spatial weight matrix is row-standardised and the time-lag of the dependent variable is included as a regressor in the model, Yu et al. (2008) argue that the spatial lag model is stable when $\rho + \beta_{NO_{2,t-1}} < 1$. This seems to be our case according to the results displayed in Table 5.4.1, although the sum of $\hat{\rho}$ and $\hat{\beta}_{NO_{2,t-1}}$ is very close to unity for model (5.2.1). The log-likelihood value is, as expected, larger for the state space model than the spatial lag model, due to the higher complexity of the former.

Figure 5.4.5 shows the Kalman filter and smoother estimates, together with the 95% confidence intervals[9] for the filter estimates, of the time-varying parameters of the spatial lag state space model (5.2.3): $\beta_{Traffic,t}$, $\alpha_{Sea,t}$ and $\alpha_{Border,t}$. The estimated time-varying effect of traffic intensity fluctuates around a positive mean. Its wide confidence intervals, together with the insignificance of $\sigma_{Traffic}$ (from Table 5.4.1), which is supposed to drive the time-variation in $\beta_{Traffic,t}$, suggests that the latter parameter is actually constant. This result implies that the non-increasing pattern of Dutch NO_2 concentrations, displayed in Figure 5.4.3a, is not due to the use of more environmentally sustainable vehicles. In fact, the constant effect of traffic intensity on pollution suggests that the partial adoption of such vehicles has not been enough to reduce their contribution to NO_2 concentrations. It can therefore be the case that other types of economic activity, such as energy consumption/production or industrial production, have become more sustainable and have therefore started offsetting, over time, the extra NO_X emissions derived by the rising number of road vehicles.

The estimated time-varying dummies, $\hat{\alpha}_{Sea,t}$ and $\hat{\alpha}_{Border,t}$, also fluctuate around a negative and positive mean, respectively. This result is in line with

[9] These confidence intervals and the forecast intervals of Figure 5.B.6 do not reflect the additional uncertainty of using estimates for the hyperparameters.

the estimates for β_{Sea} and β_{Border} discussed above. The time-variation of the unobserved components seems to be more pronounced than the one of $\beta_{Traffic,t}$, because of narrower confidence intervals. Moreover, the estimates and standard errors obtained for the hyperparameters that govern their time-variation (i.e., the respective γ and σ parameters), reported in Table 5.4.1, suggest that we can reject the hypothesis of time-constancy for both of them. Indeed, we can reject the individual hypotheses that $\gamma_{\alpha,i} = 1$ and $\sigma_{\alpha,i} = 0$, for $i = \{Sea, Border\}$, which (jointly) define a time-constant evolution of the parameter. These findings therefore suggest that the differences in NO_2 concentrations between maritime and border regions, with respect to the remaining inland regions, are time-varying[10].

The positive difference in NO_2 concentrations between regions that are located at land borders and inland regions, is, as already discussed, due to the import of pollution from neighbouring countries. The negative difference between maritime and inland regions reflects instead the fact that the former regions can export pollution to the sea, but they hardly import any pollution from the sea. Such "import/export effects" can be time-varying due, for example, to changes in meteorological factors that are not controlled for in the regressor matrix. Information about wind speed and direction is used to build the spatial weight matrix, which does not, by itself, take into account the fact that some regions are located at land borders and some other along the coast, and that therefore some of their pollution can be exported/imported abroad; the "border effect" needs indeed to be taken into account by the model in other ways, as we do. However, wind speed and direction are not also included as regressors, yet these wind characteristics can vary over time the quantity of pollution that peripheral regions import or export. For instance, the highest peak of wind speed[11], from Figure 5.4.3e, which is observed on December

[10] It should be noted that from Figure 5.4.5 it looks like the confidence intervals of the time-varying coefficients do not always deviate from zero. However, when estimated as time-constant, the spatial lag model finds the same coefficients to be significantly different from zero. A larger sample size (T) would allow us to achieve narrower confidence intervals for the state variables and to be more conclusive about this issue. This could be achieved by, for instance, employing daily rather than monthly data. However, daily data for traffic intensity are not made available by Statistics Netherlands yet.

[11] Which does not correspond to the highest peak of NO_2 concentrations, from Figure 5.4.3a,

2011, corresponds to a very low peak in both $\hat{\alpha}_{Sea,t}$ and $\hat{\alpha}_{Border,t}$. Seasonal fluctuations in inland shipping and economic activity of adjacent Belgian and German industrial areas, might also explain some of the time variation in the unobserved components. Finally, although we could not reject the hypothesis that the coefficient for $Traffic$ is time constant, Figure 5.4.5 shows that its estimate is more volatile at the beginning of the sample. Such behaviour might be explained by the implementation of legislative changes. For instance, in September 2012 the speed limit on some Dutch highways was increased from 120 km/h to 130 km/h. This may have triggered the subsequent increased volatility in the coefficient estimate for $Traffic$, which would reflect both the fact that the change was not applied to all highways, and that drivers did not adjust to it abruptly and homogeneously.

The assumptions of homoskedasticity and no serial correlation made throughout the chapter can be tested on the standardized one-step ahead forecast errors (Durbin & Koopman, 2012, Chapter 7): $\tilde{\boldsymbol{v}}_t = \boldsymbol{B}_t \boldsymbol{v}_t$, for $t = 1, \ldots, T$ with $\boldsymbol{B}_t$ such that $(\boldsymbol{F}_t)^{-1} = \boldsymbol{B}_t^{'} \boldsymbol{B}_t$. We test each series of standardized forecast errors for no serial correlation with the Ljung and Box (1978) test on 4 and 12 lags[12], and for homoskedasticity using the H(h) test described in in Durbin and Koopman (2012, Chapter 2), with h equal to 27 and 41 (which roughly correspond to a third and a half of the sample size, respectively). The p-values of the diagnostic tests are displayed in Figure 5.B.7. At the 5% significance level, the hypothesis of homoskedasticity seems to hold in general, taking also into account that we do not correct for multiple hypotheses testing here, in which case we would expect fewer rejections. The hypothesis of no serial correlation is rarely rejected by the Ljung-Box test when 12 lags are included, which suggests that the seasonality is not left over in the residuals, but captured by the meteorological regressors. However, when only 4 lags are included in the test, the hypothesis of no serial correlation is rejected much more often, indicating that there is still some time persistency that is not captured by the dynamics

which is instead observed on November 2011; it corresponds to the subsequent drop in NO_2 concentrations as they would tend to disperse with a higher wind speed.

[12] For each series, we correct the degrees of freedom of the test statistic, based on the number of parameters needed to estimate its unobserved components. See Harvey (1989, Chapter 5) for more details on the degrees of freedom correction for the Ljung-Box test.

of the state space model, and which requires further investigation.

Of wide interest it is not only to model air pollutants, but also to predict their future values based on different scenarios of economic activity. We therefore conclude this section with two forecast exercises of NO_2 for 2017 (which is the out-of-sample period in this forecast study), by considering two different scenarios of traffic intensity during the same year. In both cases, we first estimate the static parameters of models (5.2.1) and (5.2.3) based on data available until the end of 2016, and we use the so-obtained estimates in order to forecast up to twelve future monthly values of NO_2 concentrations (hence we do not adopt a moving or rolling window setting for our forecast analysis).

The forecasted values for $\boldsymbol{y}_t$, based on the state space model (5.2.3), are obtained as follows

$$\hat{\boldsymbol{y}}_{t^*+m} = (\boldsymbol{I}_n - \hat{\rho}\boldsymbol{W}_{t^*+m})^{-1}(\boldsymbol{X}_{t^*+m}\hat{\boldsymbol{\beta}}_{t^*+m} + \boldsymbol{Z}\hat{\boldsymbol{\alpha}}_{t^*+m}) \tag{5.4.3}$$

for $m = 1, \ldots, 12$ and with t^* being December 2016. In the formula above, $\hat{\rho}$ is the maximum likelihood estimate and the static elements of $\hat{\boldsymbol{\beta}}_{t^*+m}$ are the Kalman smoother estimates, of the corresponding parameters. For the time-varying parameters, we instead employ the m-step ahead Kalman filter forecasts, denoted by $\hat{\beta}_{Traffic,t^*+m}$ and $\hat{\boldsymbol{\alpha}}_{t^*+m}$, implied by their transition equations.

For the spatial lag model (5.2.1) we compute the predicted values as

$$\hat{\boldsymbol{y}}_{t^*+m} = (\boldsymbol{I}_n - \hat{\rho}\boldsymbol{W}_{t^*+m})^{-1}[\boldsymbol{X}_{t^*+m} \quad \boldsymbol{Z}]\hat{\boldsymbol{\beta}} \tag{5.4.4}$$

for $m = 1, \ldots, 12$, where now $\hat{\rho}$ and $\hat{\boldsymbol{\beta}}$ are all maximum likelihood estimates. Equations (5.4.3) and (5.4.4) also imply that we treat the spatial weight matrix and the regressors as known in the out-of-sample period. Specifically, the time lag of $\boldsymbol{y}_t$ (which is part of the regressors) is equal to $\boldsymbol{y}_{t^*}$ when $m = 1$ and $\hat{\boldsymbol{y}}_{t^*+m-1}$ when $m > 1$.

In a first forecast exercise, the values for all regressors and for the spatial weight matrix in 2017 are left unchanged from their actual observations in the same period. We are therefore in a scenario where we assume that there is no future change in traffic intensity and meteorological conditions. Figure 5.4.6a displays the time series of $\log(NO_2)$ in 2017, averaged over the COROP regions $\left(\text{i.e., } \bar{y}_{t^*+m} = \frac{1}{40}\sum_{i=1}^{40} \log(NO_{2,i,t^*+m}), \text{for } m = 1, \ldots, 12\right)$ together with the corresponding forecasts (solid lines), based on models (5.2.1) and (5.2.3), which are also averaged over the COROP regions $\left(\text{i.e., } \bar{\hat{y}}_{t^*+m} = \frac{1}{40}\sum_{i=1}^{40} \hat{y}_{i,t^*+m}, \text{for } m = 1, \ldots, 12\right)$. The forecasts obtained with both models are rather similar for the first half of the out-of-sample period, and tend to deviate more from each other afterwards. This result is attributable to the autoregressive dynamic structure for the unobserved components of the state space models, which, as the forecast horizon increases, yields predictions of such state variables equal to their zero expected value, as Figure 5.B.6 shows. This is equivalent to assuming that the "import/export effect" vanishes over time, when forecasting it, which is unrealistic and a drawback of the state space model, over the spatial lag model. The mean squared forecast error obtained with the latter model is equal to 0.113, compared to a slightly larger 0.129 yielded by the former model. Notice also that the spatial lag and the state space model tend to, respectively, under- and over-predict the level of NO_2 concentrations.

The second forecast exercise consists in reducing the level of $Traffic$ by 100%, with respect to its observed values, in all months of 2017, which is equivalent to assuming that there are no road vehicles during that year. Such a scenario is unrealistic in practice, but it allows us to understand what is the maximum decrease in NO_2 concentrations that we can expect by removing the traffic-related source of NO_X emissions. The forecasts obtained under this scenario, and with both models (5.2.1) and (5.2.3), are again displayed in Figure 5.4.6a (dashed lines). Moreover, Figure 5.4.6b shows the percentage change of the predictions of NO_2 concentrations (in levels, not in logarithms) obtained under the two scenarios[13]. We notice that the removal of traffic activ-

[13] Let $\bar{\hat{y}}^{(s_1)}_{t^*+m}$ and $\bar{\hat{y}}^{(s_2)}_{t^*+m}$ be the averaged forecasted value of $log(NO_2)$ at time $t^* + m$ under scenarios 1 and 2, respectively (with scenario 2 corresponding to a drop in $Traffic$ of 100%). The percentage change in predicted NO_2 concentrations at time $t^* + m$ is calculated as $\left(\bar{\hat{y}}^{(s_2)}_{t^*+m} - \bar{\hat{y}}^{(s_1)}_{t^*+m}\right)$ 100, since the log-difference is approximately equal to the

ity produces a gradual decrease in predicted NO_2 concentrations, which take around eights months to stabilize around their new steady level. The state space model (5.2.3) predicts a maximum percentage decrease in NO_2 concentrations of around 35%, which is in line with the TNO (2019) finding that road traffic is responsible for 31% of NO_X emissions, but it is still not consistent with the claim of Curier et al. (2014) that road vehicles are responsible for around 50% of NO_2 VCD at OMI overpass time. However, the spatial lag model (5.2.1) predicts a maximum decrease in NO_2 concentrations of around 70%, which is clearly too optimistic. This result implies that our proposed spatial lag state space model is able to provide more realistic predictions of NO_2 concentrations, under different scenarios of traffic intensity, with respect to the spatial lag model, probably due to the close-to-instability feature of the latter model discussed earlier in this section.

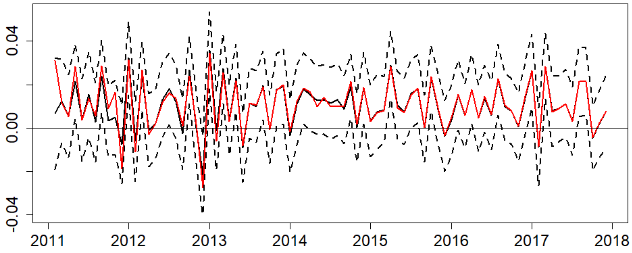
We mentioned in the Introduction that the Dutch government implemented a speed-limit reduction on highways in March 2020. The forecast analysis that we just conducted shows that we can potentially use our state space model in order to evaluate policies, such as the one just mentioned. However, we do not have data about speeds of vehicles, and it is non-trivial to quantify how many vehicles per hectare less would correspond to a maximum speed reduction of 20/30 kmph[14]. Therefore, we cannot use our model to evaluate this specific policy. However, the forecast results discussed above immediately allow us to evaluate the effect of less drastic decreases in traffic intensity, than the one we considered: e.g., a reduction in $Traffic$ by 10% in every month of 2017, is expected to yield an overall decrease in NO_2 concentrations of around 3.5%, and so on.

percentage change.

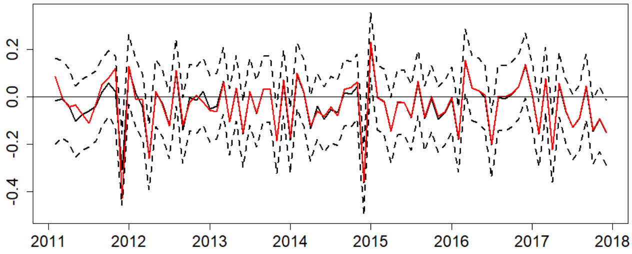
[14] There is, nonetheless, research on the effects that speed limit reductions have on air pollutants concentrations, of which Lopez-Aparicio et al. (2020) is an example.

	Spatial lag state space model (5.2.3)	Spatial lag model (5.2.1)
$\hat{\rho}$	0.5959^{***} (0.0244)	0.8828^{***} (0.0063)
$\hat{\sigma}_{\varepsilon}$	0.2056^{***} (0.0450)	0.1342^{***} (0.0134)
$\hat{\beta}_{Intercept}$	1.0767 (0.0888)	0.2271^{***} (0.0595)
$\hat{\beta}_{NO_{2,t-1}}$	0.2570 (0.0124)	0.0815^{***} (0.0076)
$\hat{\beta}_{Traffic}$		0.0096^{***} (0.0007)
$\hat{\beta}_{Temp}$	0.0120 (0.0022)	0.0031^{***} (0.0008)
$\hat{\beta}_{Sun}$	-0.0013 (0.0002)	-0.0003^{***} (0.00006)
$\hat{\beta}_{Rain}$	-0.0012 (0.0002)	-0.0004^{***} (0.00008)
$\hat{\beta}_{Sea}$		-0.0433^{***} (0.0059)
$\hat{\beta}_{Border}$		0.0440^{***} (0.0060)
$\hat{\gamma}_{\alpha,Sea}$	-0.2781^{***} (0.0385)	
$\hat{\gamma}_{\alpha,Border}$	0.5956^{***} (0.0552)	
$\hat{\sigma}_{Traffic}$	0.0930 (0.3433)	
$\hat{\sigma}_{\alpha,Sea}$	0.4679^{***} (0.0453)	
$\hat{\sigma}_{\alpha,Border}$	1.6867^{***} (0.0091)	
Log-likelihood	-441.895	-1420.94

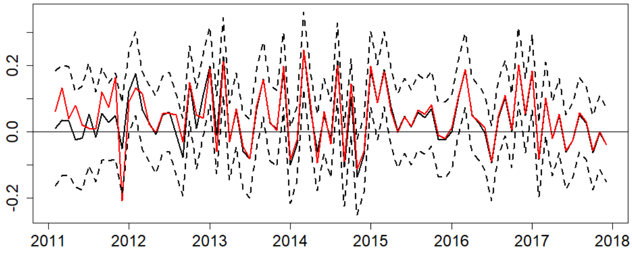
Table 5.4.1: In-sample results. Maximum likelihood and Kalman smoother estimates of the models' static parameters. The standard errors are reported in parentheses; *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01, from a Wald χ^2 test for individual hypotheses (when possible, i.e., not for the static β coefficients that are estimated by Kalman smoothing). Harvey (1989, Chapter 5) explains how to modify the test when the null hypothesis corresponds to the boundary of the parameter space, which is the case for the standard deviations. Namely, the p-values from the Wald test statistic have to be compared to a critical level of 2α, instead of α, because under the null hypothesis its asymptotic distribution is not χ^2_1 as usual any more.



(a) $\hat{\beta}_{Traffic,t}$

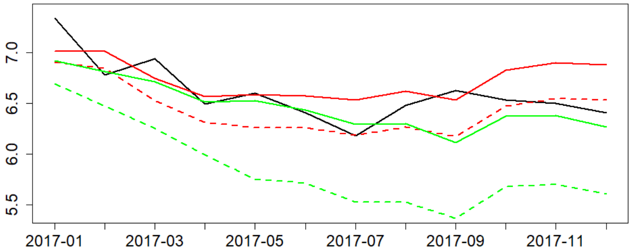


(b) $\hat{\alpha}_{Sea,t}$

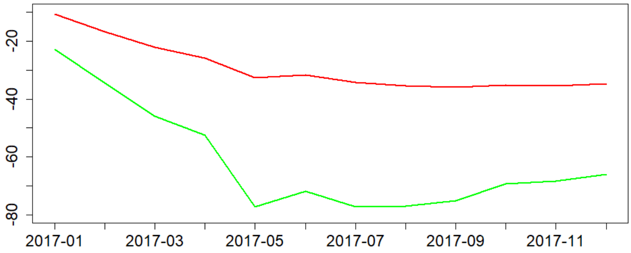


(c) $\hat{\alpha}_{Border,t}$

Figure 5.4.5: Kalman filter (black solid lines) and Kalman smoother (red solid lines) estimates of $\beta_{Traffic,t}$, $\alpha_{Sea,t}$ and $\alpha_{Border,t}$. The dashed black lines are the 95% confidence intervals of the Kalman filter estimates.



(a) Forecasts



(b) Percentage changes in predicted NO_2 concentrations

Figure 5.4.6: In Figure 5.4.6a, the black line displays the observed values of $\log(NO_2)$ (averaged over the Dutch COROP regions). The solid green and red lines represent the average (over the Dutch COROP regions) of the forecasted values for $\log(NO_2)$ based on, respectively, the spatial lag model (5.2.1) and the state space model (5.2.3), when $Traffic$ is left unchanged. The dashed lines refer to the same forecasts obtained when $Traffic$ is decreased by 100%. Figure 5.4.6b displays the percentage changes in the predictions of NO_2 concentrations obtained under the scenarios that $Traffic$ is left unchanged and $Traffic$ is decreased by 100%. The green and red lines refer to the results obtained with models (5.2.1) and (5.2.3), respectively.

5.5 Conclusions

In this chapter we propose a new dynamic multivariate econometric model, called spatial lag state space model, which adds, to standard state space models, a spatial (cross-sectional) dependence in the dependent variable. Such a model has the advantage, over spatial lag models, of allowing for time-varying regression coefficients and latent variables that can capture, for instance, inertial dynamics or time-varying differences between the intercepts of the dependent variable. A Monte Carlo simulation study shows that all of these state variables, i.e., (time-varying) regression coefficients and unobserved components, as well as the remaining static parameters of the model, can be efficiently and accurately estimated by Kalman filtering/smoothing and maximum likelihood, respectively.

We apply our proposed model to a regional analysis of nitrogen dioxide (NO_2) concentrations in the Netherlands, over the period January 2011-December 2017. We employ monthly data, and traffic intensity and meteorological variables as predictors of the pollutant of interest. The spatial weight matrix, which defines the connectivity of the geographical regions, is based on data about wind speed and direction in order to capture the spatial spillovers of NO_2 concentrations, which are due to its transportation by the wind.

Our empirical results suggest that the spatial dependence among Dutch regions, in terms of NO_2 concentrations, is positive and strong, thus supporting the choice of a spatial type of model. On top of the positive impact of road traffic, temperature and the time lag of the pollutant's concentrations, we find sunlight and rain to have a negative effect on NO_2 concentrations.

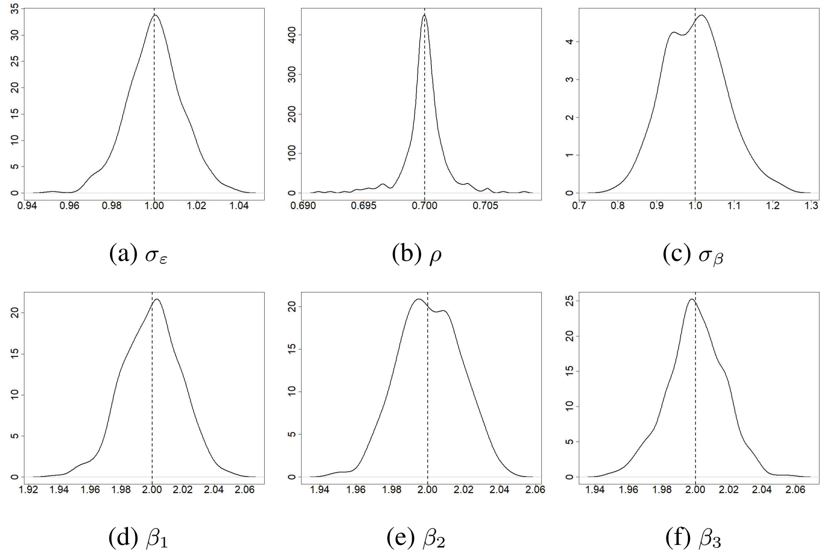
Despite an observed overall increase in Dutch traffic intensity, over the sample considered, we do not see a similar rise in NO_2 concentrations. However, we do not find the effect of traffic intensity on the latter variable to be time-varying, suggesting that other economic activities (that we do not control for in our model yet) have been subject to a sustainable change which decreased their contributions to the quantity of tropospheric NO_2. We employ unobserved components in order to model the so-called "border effect", i.e., to take into account the fact that NO_2 concentrations of peripheral Dutch regions can affect (and be affected by) neighbouring regions that are not part of the spatial units under study. We find time-varying differences, in the levels of tropospheric NO_2 concentration, between peripheral regions with respect to inland regions, thus supporting the use of a state space model in order to account for such time-changing features. Specifically, the pollution of maritime regions is generally lower than inland regions, due to the fact that they export more NO_2 to the sea, than they import from the sea. On the contrary, regions bordering with Belgium and/or Germany tend to import more NO_2 from these neighbouring countries than they export, resulting in an overall higher level of pollution, with respect to inland regions. These differences are likely time-varying due to meteorological conditions and seasonal fluctuations of economic activities abroad.

The proposed spatial lag state space model is further employed to forecast, up to twelve-months-ahead, regional NO_2 concentrations, under different scenarios of traffic intensity. When leaving the level of road traffic unchanged in the forecast period, with respect to its actual observations, our model yields a similar performance compared to a more static spatial lag model. However, when hypothesising a complete removal of road traffic in the forecast period, our state space approach yields much more realistic forecasts of NO_2 concentrations, predicting an overall decrease of 35$\%$, which is consistent with the TNO (2019) finding that vehicles are responsible for 31$\%$ of NO_X emissions. The spatial lag model, instead, too optimistically predicts a decrease in pollution of about 70$\%$. With both models, it is estimated to take around eight months for the concentrations of NO_2 to achieve their new steady level, after a given decrease in traffic intensity in the forecasted period.

In the Introduction we mentioned that NO_X is responsible for the creation of stratospheric ozone and secondary aerosols, which cool the Earth's surface. There is therefore the possibility of an indirect effect of NO_2 concentrations on temperature, which is not currently taken into account in the econometric model specification. In case of such contemporaneous and reverse effect, the maximum likelihood estimators would suffer from a simultaneity bias. This problem could be tackled by employing a simultaneous equations structure for the state space model, thus augmenting the observed dependent vector with temperature, as well as aerosol and stratospheric ozone concentrations variables. However, in a spatial econometric setting this would dramatically increase the dimensionality of the dependent variable, hence hampering a feasible estimation of the model. Econometric models that do not include many cross-sectional observations would be more suited to study this type of chemical interactions and reverse effects. Such a model is, for instance, employed by Montamat and Stock (2020) who take an instrumental variables approach in order to deal with the simultaneity bias due to the direct effect that carbon dioxide, as a greenhouse gas, has on temperature, and viceversa.

5.A Monte Carlo simulation results

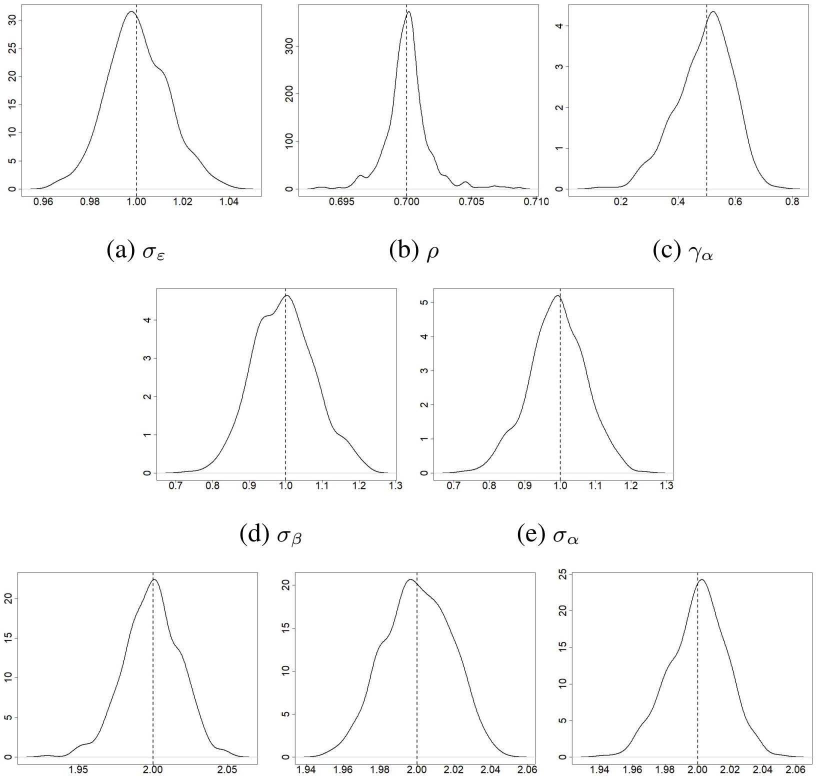
5.A.1 $T = 80$



(a) σ_ε (b) ρ (c) σ_β

(d) β_1 (e) β_2 (f) β_3

Figure 5.A.1: Distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, for model specification 1 and with $T = 80$. The β coefficients are estimated by the Kalman smoother, and the remaining parameters (the hyperparameters) by maximum likelihood. The dashed lines represent the true values of the parameters.

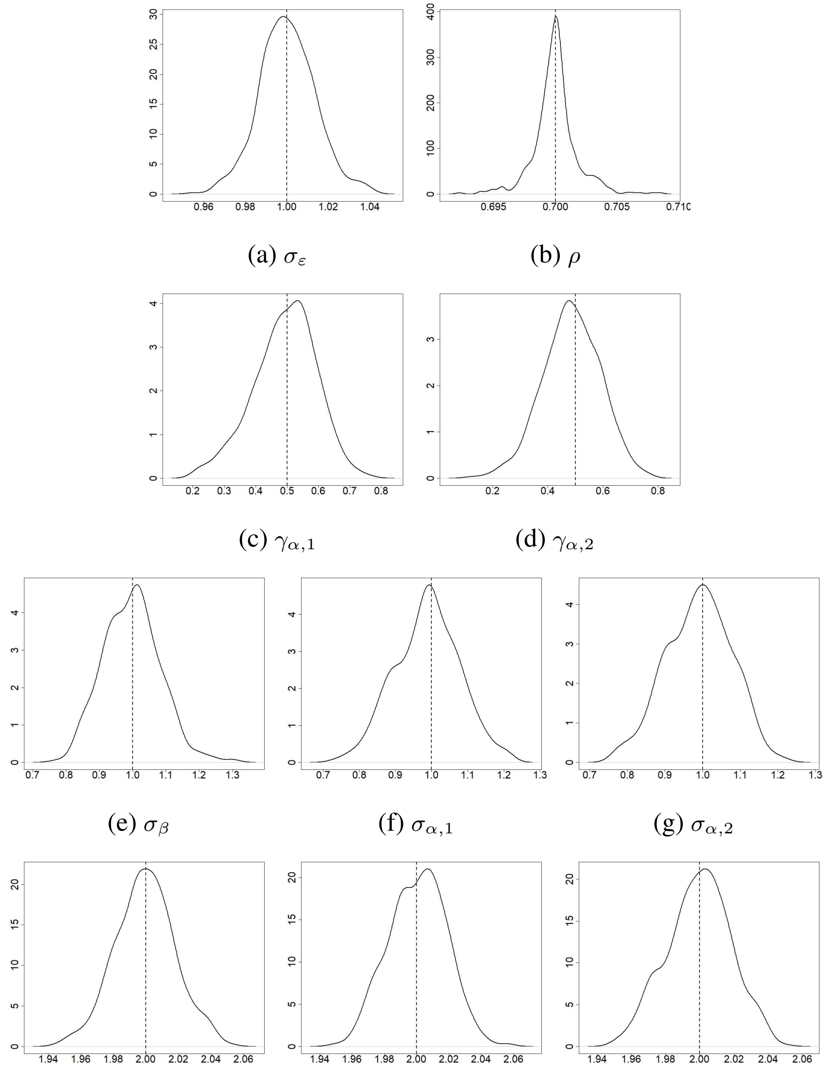


(a) σ_ε (b) ρ (c) γ_α

(d) σ_β (e) σ_α

(f) β_1 (g) β_2 (h) β_3

Figure 5.A.2: Distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, for model specification 2 and with $T = 80$. The β coefficients are estimated by the Kalman smoother, and the remaining parameters (the hyperparameters) by maximum likelihood. The dashed lines represent the true values of the parameters.



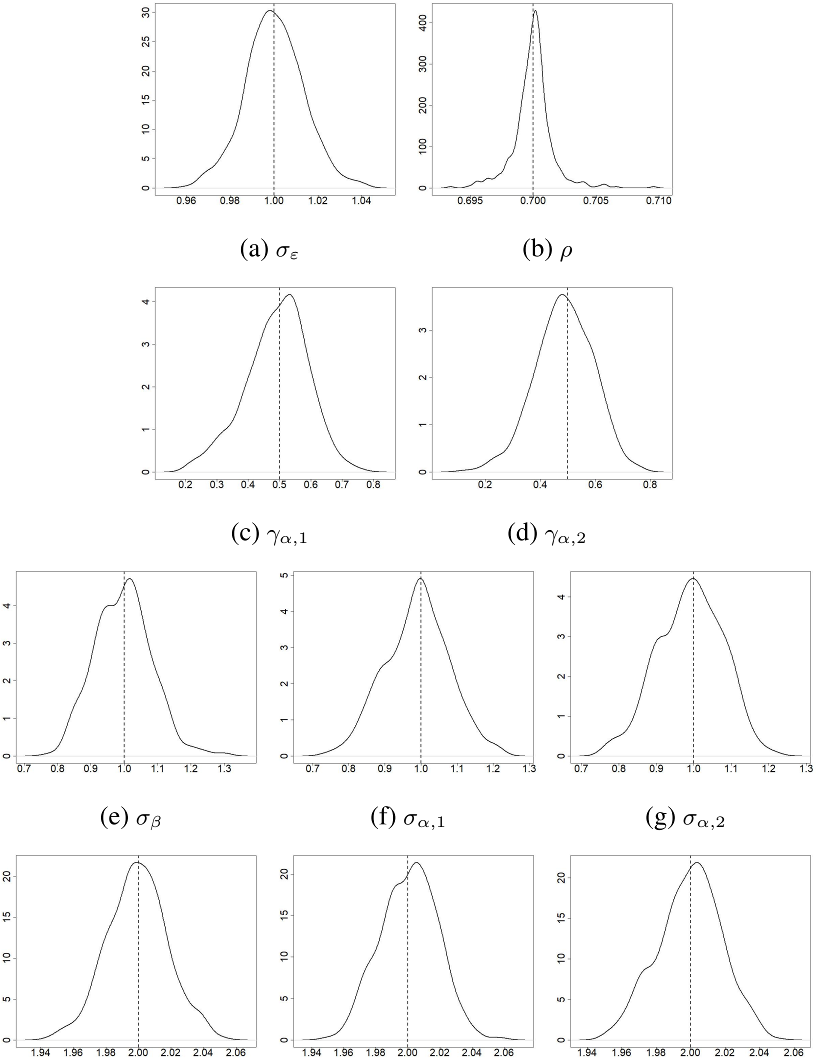
(a) σ_ε (b) ρ

(c) $\gamma_{\alpha,1}$ (d) $\gamma_{\alpha,2}$

(e) σ_β (f) $\sigma_{\alpha,1}$ (g) $\sigma_{\alpha,2}$

(h) β_1 (i) β_2 (j) β_3

Figure 5.A.3: Distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, for model specification 3 and with $T = 80$. The β coefficients are estimated by the Kalman smoother, and the remaining parameters (the hyperparameters) by maximum likelihood. The dashed lines represent the true values of the parameters.



(a) σ_ε

(b) ρ

(c) $\gamma_{\alpha,1}$

(d) $\gamma_{\alpha,2}$

(e) σ_β

(f) $\sigma_{\alpha,1}$

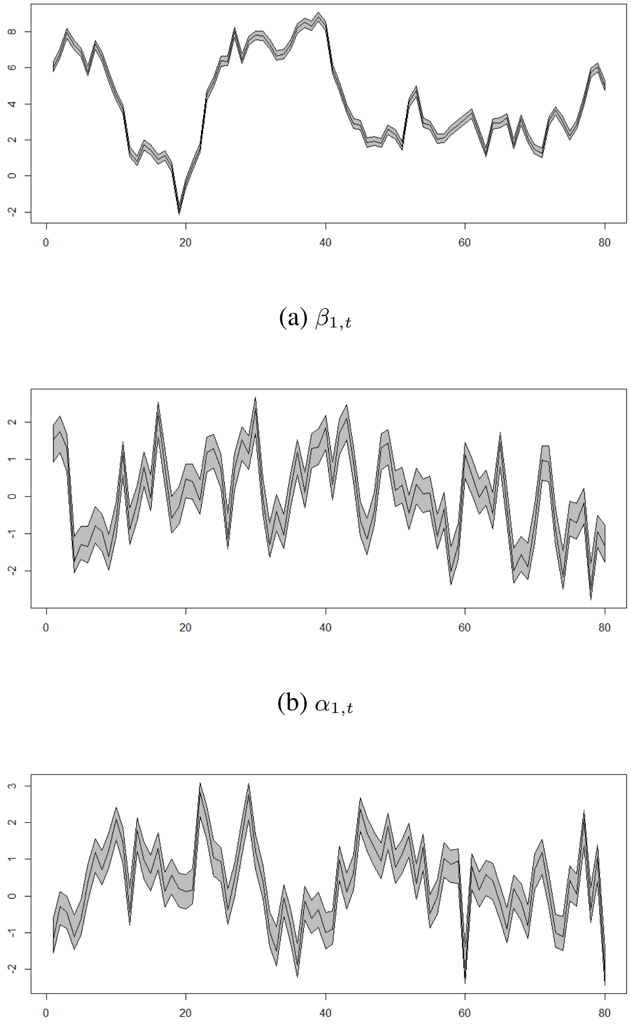
(g) $\sigma_{\alpha,2}$

(h) β_1

(i) β_2

(j) β_3

Figure 5.A.4: Distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, for model specification 4 and with $T = 80$. The β coefficients are estimated by the Kalman smoother, and the remaining parameters (the hyperparameters) by maximum likelihood. The dashed lines represent the true values of the parameters.

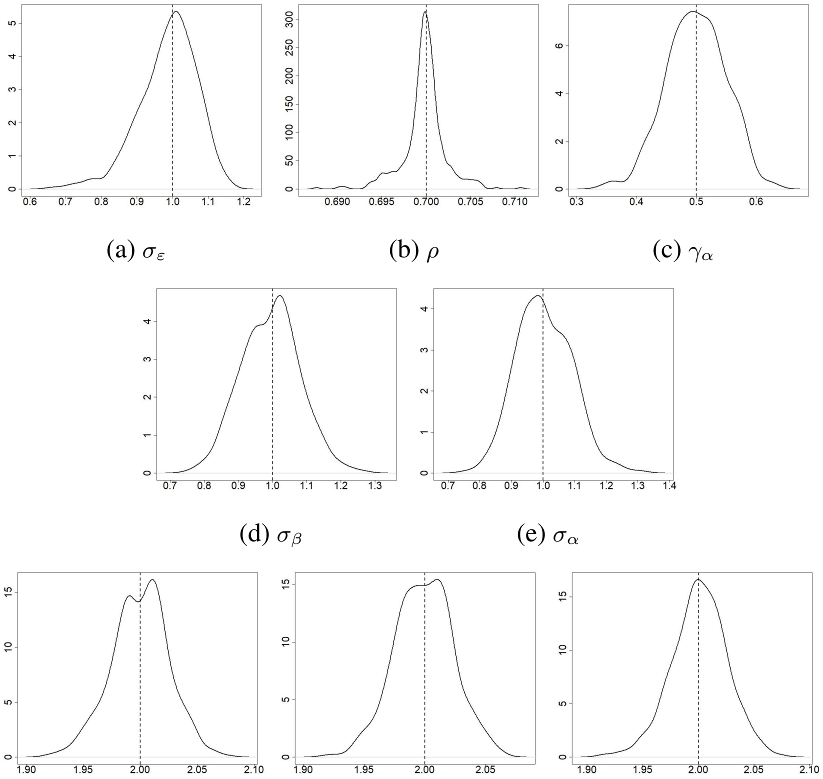


(a) $\beta_{1,t}$

(b) $\alpha_{1,t}$

(c) $\alpha_{2,t}$

Figure 5.A.5: True $\beta_{1,t}$, $\alpha_{1,t}$ and $\alpha_{2,t}$ (black lines) and respective 5% and 95% percentiles of the Kalman smoother estimates (gray shaded areas), based on the Monte Carlo replicates. The results refer to model specification 4 with $T = 80$.



(a) σ_ε

(b) ρ

(c) γ_α

(d) σ_β

(e) σ_α

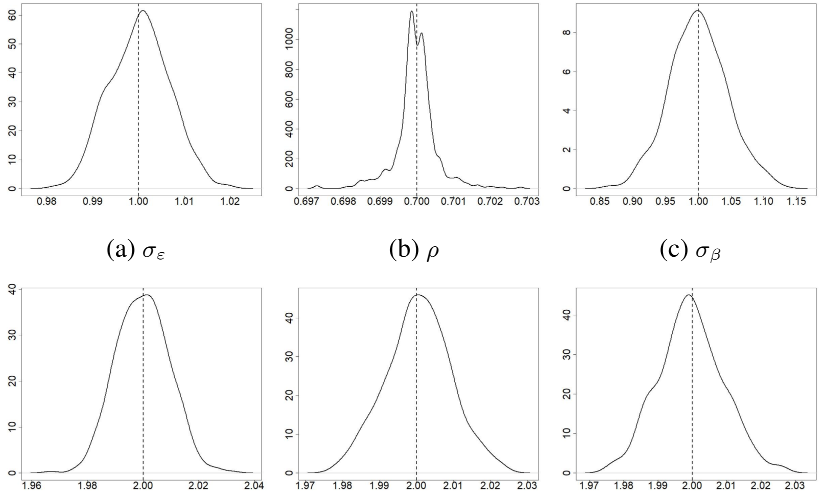
(f) β_1

(g) β_2

(h) β_3

Figure 5.A.6: Distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, for model specification 5 and with $T = 80$. The β coefficients are estimated by the Kalman smoother, and the remaining parameters (the hyperparameters) by maximum likelihood. The dashed lines represent the true values of the parameters.

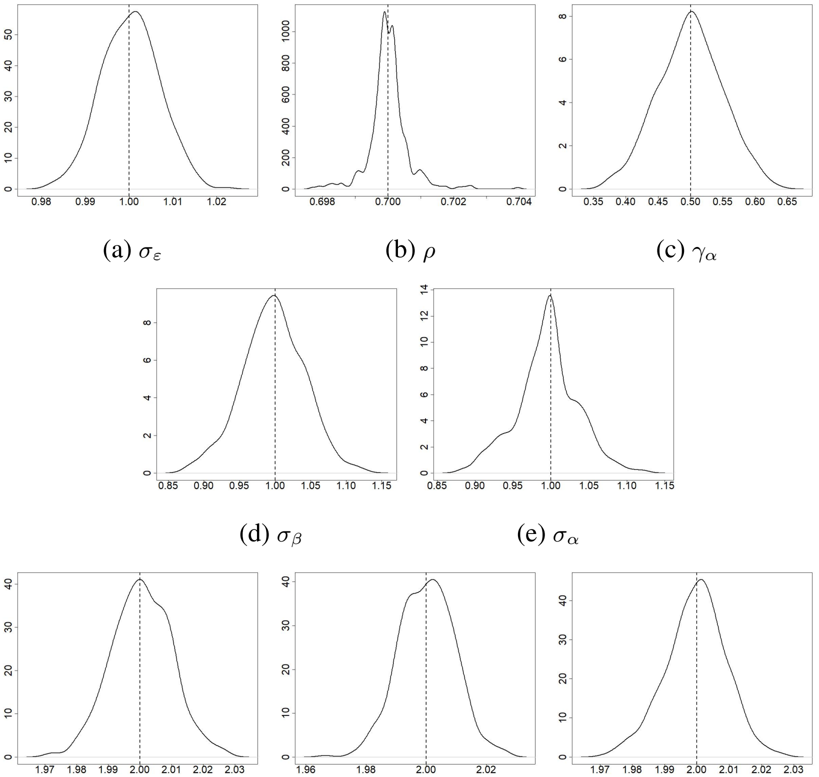
5.A.2 $T = 300$



(a) σ_ε (b) ρ (c) σ_β

(d) β_1 (e) β_2 (f) β_3

Figure 5.A.7: Distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, for model specification 1 and with $T = 300$. The β coefficients are estimated by the Kalman smoother, and the remaining parameters (the hyperparameters) by maximum likelihood. The dashed lines represent the true values of the parameters.



(a) σ_ε

(b) ρ

(c) γ_α

(d) σ_β

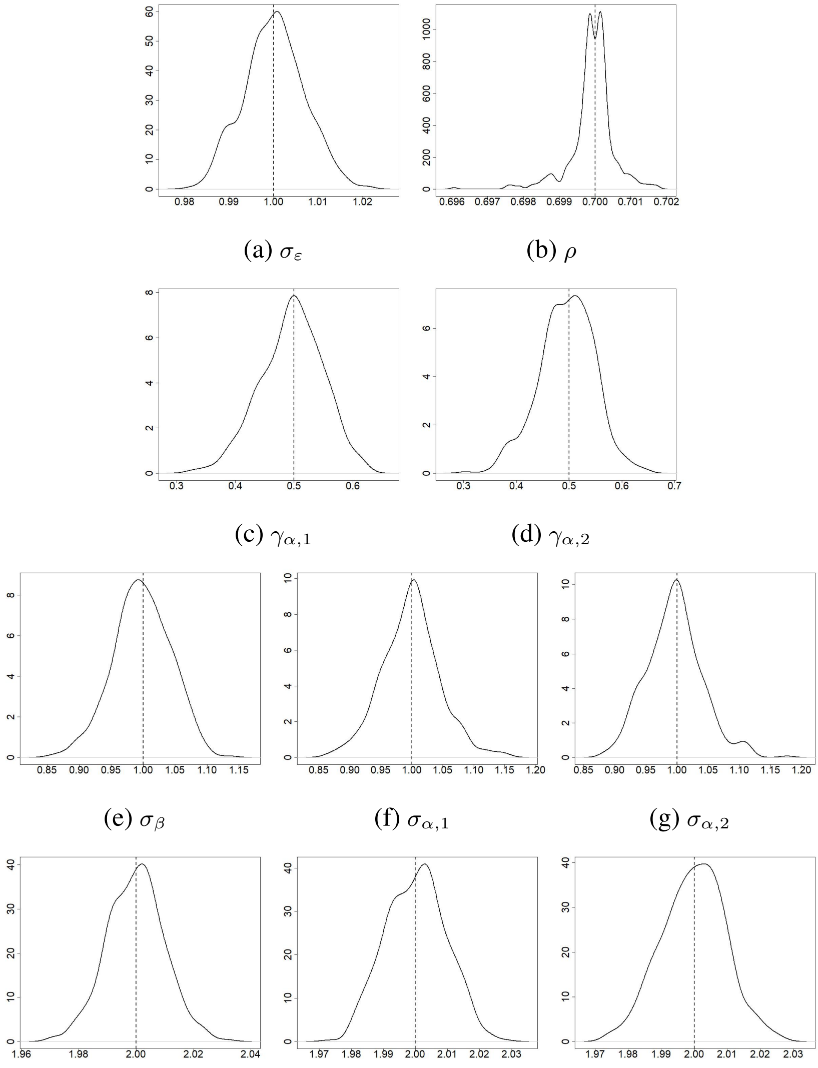
(e) σ_α

(f) β_1

(g) β_2

(h) β_3

Figure 5.A.8: Distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, for model specification 2 and with $T = 300$. The β coefficients are estimated by the Kalman smoother, and the remaining parameters (the hyperparameters) by maximum likelihood. The dashed lines represent the true values of the parameters.



(a) σ_{ε}

(b) ρ

(c) $\gamma_{\alpha,1}$

(d) $\gamma_{\alpha,2}$

(e) σ_{β}

(f) $\sigma_{\alpha,1}$

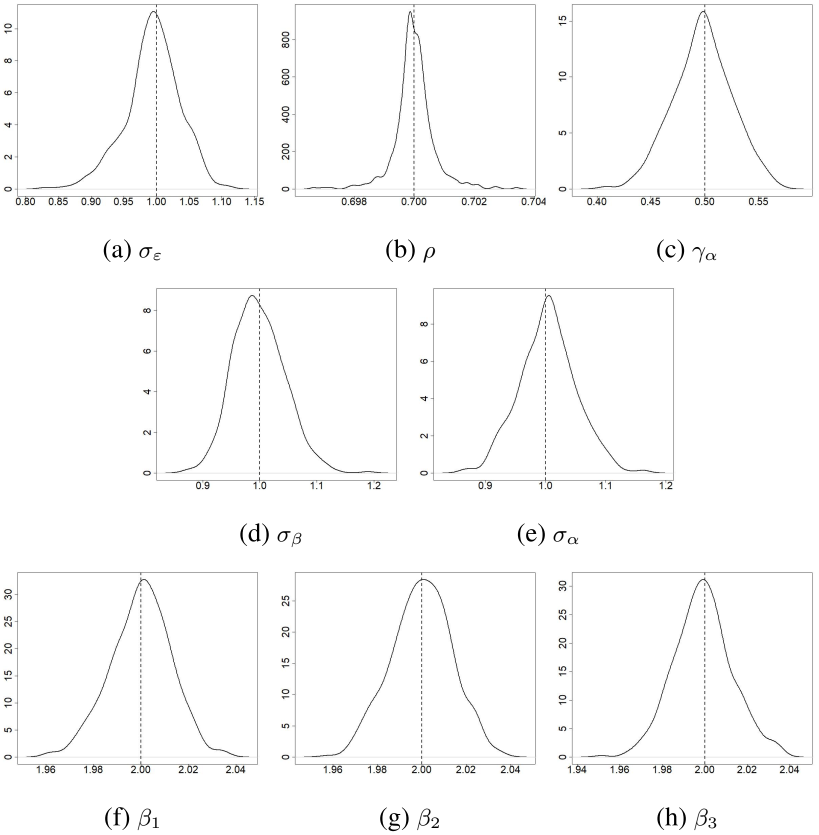
(g) $\sigma_{\alpha,2}$

(h) β_1

(i) β_2

(j) β_3

Figure 5.A.9: Distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, for model specification 3 and with $T = 300$. The β coefficients are estimated by the Kalman smoother, and the remaining parameters (the hyperparameters) by maximum likelihood. The dashed lines represent the true values of the parameters.



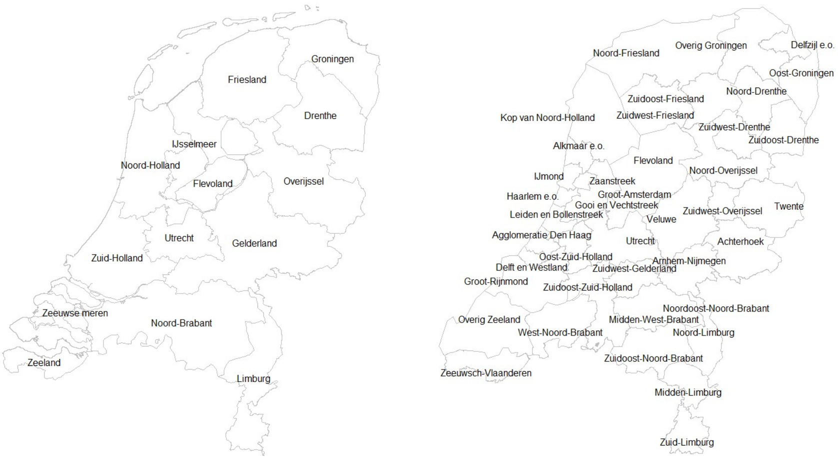
(a) σ_ε (b) ρ (c) γ_α

(d) σ_β (e) σ_α

(f) β_1 (g) β_2 (h) β_3

Figure 5.A.10: Distributions of the maximum likelihood and Kalman smoother estimators of the static parameters, based on the Monte Carlo replicates, for model specification 5 and with $T = 300$. The β coefficients are estimated by the Kalman smoother, and the remaining parameters (the hyperparameters) by maximum likelihood. The dashed lines represent the true values of the parameters.

5.B Data characteristics and additional empirical results



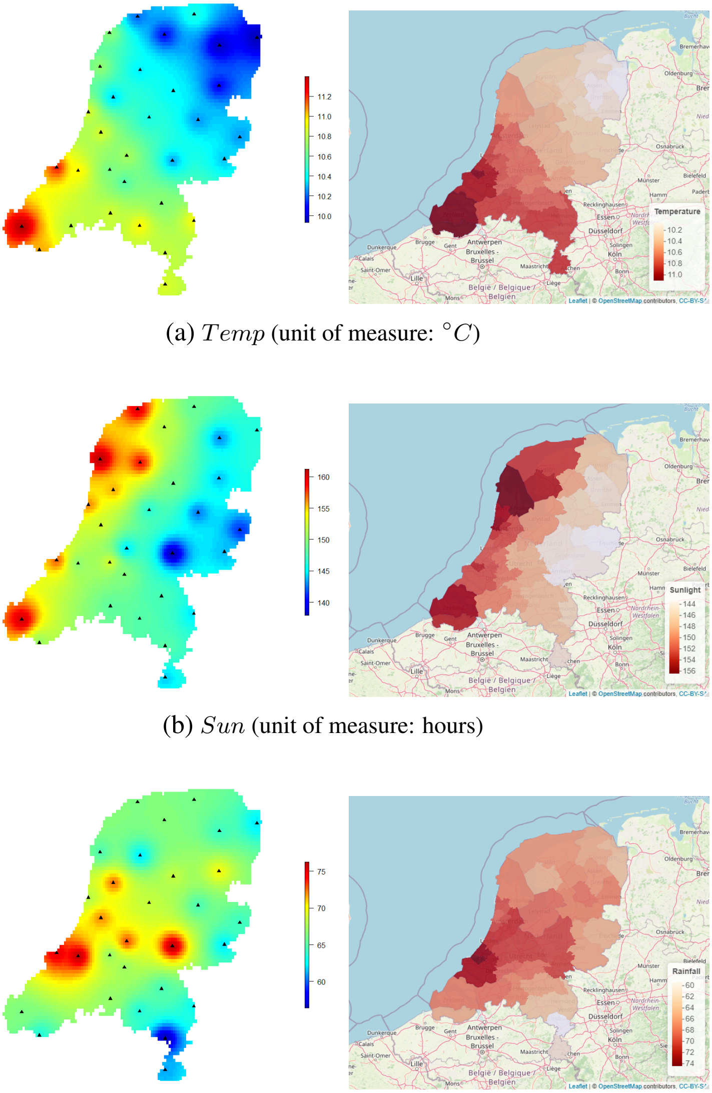
(a) The 14 Dutch provinces.

(b) The 40 Dutch COROP regions.

Figure 5.B.1: Maps of the provinces and the COROP regions of the Netherlands.

Variable	Description	Units	Mean	St. dev.	Min	Max
NO_2	$\mathrm{NO_2}$ concentrations	10^{13} molec./cm^2	889.17	365.21	224.18	2604.33
$Traffic$	road vehicles	number of vehicles/hectare	3.08	3.7	0.22	16.7
$Temp$	temperature	$°C$	10.59	5.31	-0.2	19.82
Sun	sunlight duration	hours	149.34	65.1	19.24	278.88
$Rain$	amount of rainfall	mm	67.49	36.15	3.8	193.9
$WindSpd$	wind speed	m/s	4.79	1.15	2.35	10.15
$WindDir$	wind direction	degrees	234.47	60.2	0.69	359.88

Table 5.B.1: Descriptive statistics for tropospheric NO_2 concentrations, traffic intensity and the meteorological variables. Sunshine and rainfall variables are equal to zero when failing to reach 0.05 hours and 0.05 mm per day, respectively. The descriptive statistics refer to the monthly variables (so they have to be interpreted as "per month").

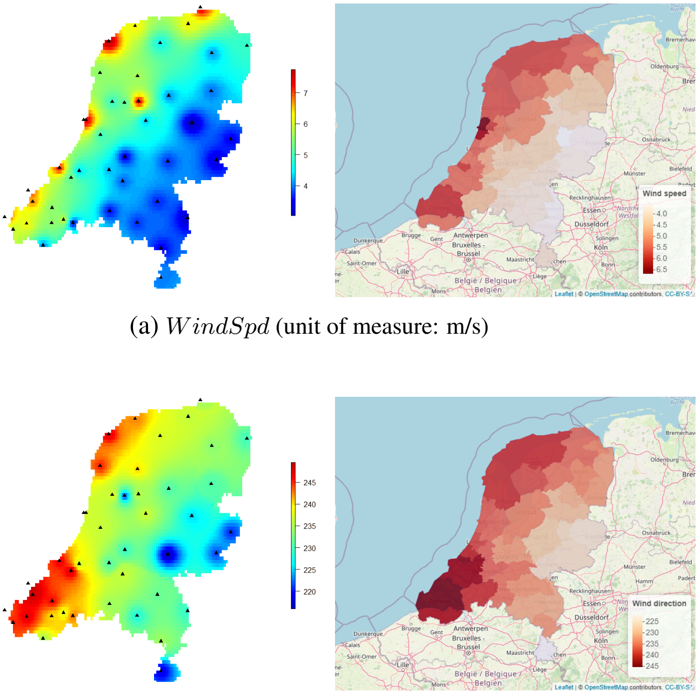


(a) $Temp$ (unit of measure: $^{\circ}C$)

(b) Sun (unit of measure: hours)

(c) $Rain$ (unit of measure: mm)

Figure 5.B.2: The meteorological variable $Temp$ has been observed by 32 measurement stations, whereas Sun and $Rain$ by 30 measurement stations. The left panels show the interpolated values of the variables, averaged over the months, together with the location of the measurement stations (black triangles). The right panels display the transformed variables into COROP regional data, averaged over the months.



(a) $WindSpd$ (unit of measure: m/s)

(b) $WindDir$ (unit of measure: degrees)

Figure 5.B.3: $WindSpd$ and $WindDir$ have both been observed by 44 measurement stations. The left panels show the interpolated values of the variables, averaged over the months, and the location of the measurement stations (black triangles). The right panels display the transformed variables into COROP regional data, averaged over the months.

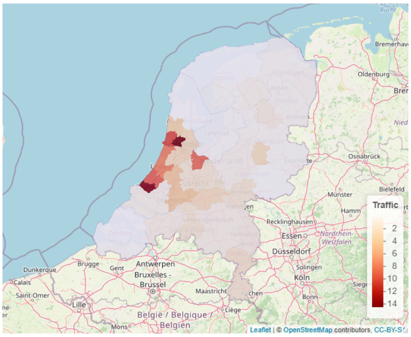


Figure 5.B.4: Spatial visualisation of traffic intensity ($Traffic$), averaged over the months, for the COROP regions. Unit of measure: number of vehicles/hectare.

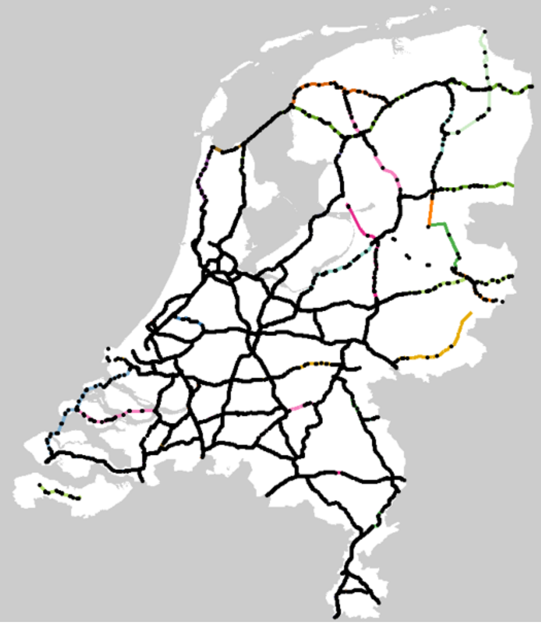
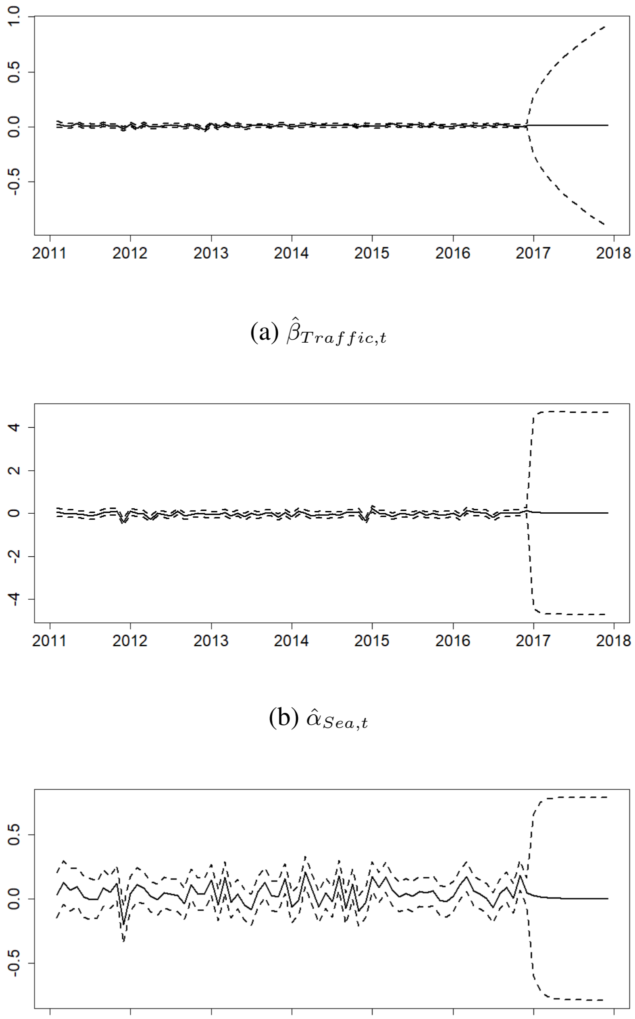


Figure 5.B.5: Location of road sensors on the Dutch highways. Source: Statistics Netherlands.

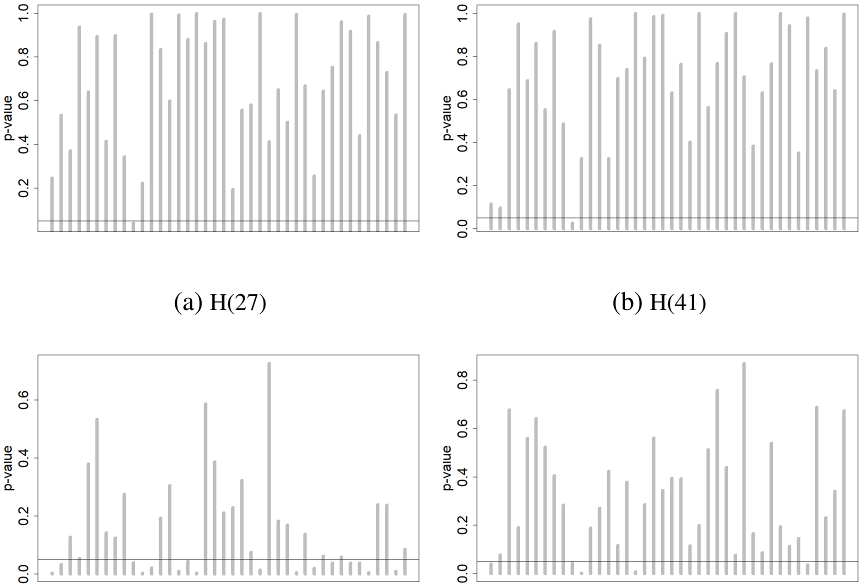


(a) $\hat{\beta}_{Traffic,t}$

(b) $\hat{\alpha}_{Sea,t}$

(c) $\hat{\alpha}_{Border,t}$

Figure 5.B.6: Kalman filter (solid lines) estimates (until 2016) and forecasts (in 2017) of $\beta_{Traffic,t}$, $\alpha_{Sea,t}$ and $\alpha_{Border,t}$. The dashed lines are the 95% confidence intervals of the Kalman filter estimates and forecasts.



(a) H(27)

(b) H(41)

(c) LB(4)

(d) LB(12)

Figure 5.B.7: P-values of the diagnostic tests applied to the standardized forecast errors. H(h) is the test for homoskedasticity, with h equal to 27 and 41. LB(q) is the Ljung-Box test for no serial correlation with q equal to 4 and 16 lags. The horizontal line represents the 0.05 significance level.

The shape files for Europe, the Dutch provinces and the Dutch COROP regions have been downloaded, respectively, from https://tapiquen-sig.jimdo.com, https://gadm.org and https://www.imergis.nl.

6 Summary and general discussion

This thesis explores how state space models, which are a type of econometric models designed to analyse time series data, can be employed to achieve more accurate and realistic estimates of official statistics, and to model and forecast regional concentrations of air pollutants. Specifically, a novel approach is presented, which incorporates survey-based, claimant counts and Google Trends data in order to provide more timely and accurate estimates of Dutch unemployment, over only employing survey-based data. A new method is proposed to model the relationship between the latter and claimant counts data as time-varying, which allows us to promptly tackle changes in such relationship and therefore achieve more realistic real-time estimates of Dutch unemployment. Time-varying relationships can potentially be modelled with other, already existing, econometric techniques, than the one proposed in this thesis, and the reasons why they have not been considered further are here documented. Finally, a novel spatial type of state space model is employed in order to model regional concentrations of nitrogen dioxide (NO_2) in the Netherlands. The (time-varying) effects on this air pollutant of meteorological conditions, traffic intensity and geographical location of the Dutch regions, are accounted for in the model. The latter is further used to forecast regional NO_2 concentrations for different scenarios of traffic intensity, and can therefore be potentially employed for evaluation of pollution-reduction policies. This chapter is continued by outlining the main conclusions of the thesis in more detail.

In Chapter 2 we propose a new methodological approach to include a high-dimensional, and available in real time, auxiliary series in state space models. A Monte Carlo simulation study shows that the nowcast accuracy of state variables of interest (obtained with the Kalman filter) can so be improved. Empirically the high-dimensional auxiliary series is represented by Google Trends of search terms related to economic uncertainty and job-search activity. We find that the series do improve the estimation and nowcast accuracy of Dutch unemployment, over only using survey-based data. Moreover, if we also include claimant count data, such improvement can be further boosted. The magnitude of these gains is however sensitive to several aspects: it depends on whether we employ monthly or weekly Google Trends, and whether and how the high-dimensional econometric techniques (penalised regressions and factor models) are implemented in order to filter out the noise of Google Trends. Nevertheless, an important finding is that in the worse case scenario, that is when Google Trends do not show any predictive power in estimating/nowacasting Dutch unemployment, our method is able to ignore the information coming from these auxiliary series, thus yielding similar results as when only survey-based data are employed.

Take-home message from Chapter 2: the more data sources are included in the state space model, the better we can estimate and nowcast the Dutch unemployment, but be careful with the way you handle the noise of Google Trends. If the latter series turn out to not be related to the unemployment, do not worry because the model will automatically ignore them.

Chapter 3 focuses on modelling the relationship between survey-based data and claimant counts data as time-varying, in the state space model mentioned above. We do so with two methods: cubic splines, and a new approach based on indirect inference and particle filtering, which still makes use of cubic splines. According to a Monte Carlo simulation study, both methods are able to properly estimate the time-varying relationship (represented by a time-varying state correlation), as well as the remaining parameters and state variables of the model. The newly proposed methodology is affected by a strong volatility in estimating the time-varying correlation, contrary to the cubic splines and a time-constant estimation of the correlation, but it outperforms the latter in real-time estimation of the time-varying relationship and state variables of interest. This implies that the approach based on indirect inference and particle filtering is able to promptly tackle changes in the time-varying relationship, which in turn yields realistic real-time estimates of Dutch unemployment. We find that the magnitude of the relationship between survey-based and claimant count data, otherwise very strong, weakened after the financial crisis of 2008 and, to a lesser extent, due to a legislative change in claimant count data collection in 2015. Our method could potentially be extended to modelling a time-varying relationship with Google Trends data as well, but its computationally intensive feature makes it unsuited to deal with high-dimensional models.

Take-home message from Chapter 3: if you are interested in realistic real-time estimates of Dutch unemployment, while using survey-based and claimant count data, then model the relationship between these data sources as time-varying with our proposed method. Be aware that events like economic crises and legislative changes in data collection can likely vary such relationship over time. Also, make sure to bring a book to read while you wait for the results!

In Chapter 4 we tried to use other, already existing, methodologies in order to model the time-varying relationship discussed above: a score driven approach, the Extended Kalman filter and the importance sampling method. However, the former approach did not yield a satisfactory performance in estimating the time-varying relationship, based on a Monte Carlo simulation study. Moreover, we did not find a way to feasibly implement the latter two methods, which are challenged by our specific type of nonlinear state space model. We anyway opted for documenting our attempts in this thesis.

Take-home message from Chapter 4: if you are struggling to solve a research puzzle, I feel for you.

Chapter 5 evolves around the development of a state space model suited for analysing and forecasting regional concentrations of nitrogen dioxide (NO_2) in the Netherlands. Our proposed state space model has the novel feature of incorporating a spatial structure that is often crucial to study environmental variables, such as our air pollutant of interest. A Monte Carlo simulation study illustrates that this spatial state space model is accurately estimated by Kalman filtering and maximum likelihood. In our empirical study we find NO_2 to be significantly affected by meteorological conditions and to be subject to a strong spatial dependence, explained by the role of the wind in spatially transporting air pollution. According to our results, the effect of traffic intensity on NO_2 concentrations is positive but not changing over time, which suggests that the adoption of environmentally sustainable vehicles has not been strong enough to have a pollution-reduction effect. However, differences between peripheral and inland regions are found to be time-varying, likely due to changes in meteorological factors and in economic activities abroad. In general, maritime regions tend to have less pollution than inland regions, because they can export NO_2 to the sea and hardly import it from abroad. The opposite holds for regions located at land borders as they import pollution from the neighbouring countries of Belgium and Germany. We finally employ our state space model to forecast regional NO_2 concentrations following an hypothetical 100% reduction in traffic intensity. We find an overall realistic reduction of 35% in NO_2 concentrations, with respect to their observed values, and that it takes around eight months in order for NO_2 concentrations to achieve their new steady level.

Take-home message from Chapter 5: A spatial approach is often needed for modelling environmental variables, such as NO_2 concentrations. Moreover, a state space model allows us to discover time-varying features of these concentrations, such as time-changing border effects. The adoption of environmentally sustainable vehicles has not been strong enough in order to significantly reduce NO_2 concentrations, but the latter can be definitely, and relatively quickly, cut down by decreasing the number of polluting vehicles.

Bibliography

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Springer.

Anselin, L. and J. Le Gallo (2006). “Interpolation of Air Quality Measures in Hedonic House Price Models: Spatial Aspects”. In: *Spatial Economic Analysis* 1.1, pp. 31–52.

Antolin-Diaz, J., T. Drechsel, and I. Petrella (2017). “Tracking the Slowdown in Long-run GDP Growth”. In: *Review of Economics and Statistics* 99.2, pp. 343–356.

Askitas, N. and K. F. Zimmermann (2009). “Google Econometrics and Unemployment Forecasting”. In: *Applied Economics Quarterly* 55.2, pp. 107–120.

Bai, J. (2004). “Estimating Cross-section Common Stochastic Trends in Nonstationary Panel Data”. In: *Journal of Econometrics* 122.1, pp. 137–183.

Bai, J. and S. Ng (2002). “Determining the Number of Factors in Approximate Factor Models”. In: *Econometrica* 70.1, pp. 191–221.

— (2008). “Forecasting Economic Time Series Using Targeted Predictors”. In: *Journal of Econometrics* 146.2, pp. 304–317.

Bailar, B. (1975). “The Effects of Rotation Group Bias on Estimates from Panel Surveys”. In: *Journal of the American Statistical Association* 70.349, pp. 23–30.

Bańbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013). *Now-casting and the Real-time Data Flow*. Working Paper Series 1564. European Central Bank.

Barigozzi, M. and M. Luciani (2017). *Common Factors, Trends, and Cycles in Large Datasets*. Finance and Economics Discussion Series 2017-111. Board of Governors of the Federal Reserve System (U.S.)

Bennedsen, M., E. Hillebrand, and S. J. Koopman (2019). “Trend Analysis of the Airborne Fraction and Sink Rate of Anthropogenically Released CO2”. In: *Biogeosciences* 16.18, pp. 3651–3663.

— (2021). “Modeling, Forecasting, and Nowcasting U.S. CO2 Emissions Using Many Macroeconomic Predictors”. In: *Energy Economics* 96.

Blasques, F., S. J. Koopman, K. Łasak, and A. Lucas (2016). "In-sample Confidence Bands and Out-of-Sample Forecast Bands for Time-varying Parameters in Observation-driven Models". In: *International Journal of Forecasting* 32.3, pp. 875–887.

Bocci, C. and A. Petrucci (2016). "Spatial Information and Geoadditive Small Area Models". In: *Analysis of Poverty Data by Small Area Estimation*. Ed. by M. Pratesi. Wiley Series in Survey Methodology. Wiley. Chap. 12, pp. 245–259.

Boersma, K. F., H. J. Eskes, R. J. Dirksen, R. J. van der A, J. P. Veefkind, P. Stammes, V. Huijnen, Q. L. Kleipool, M. Sneep, J. Claas, J. Leitão, A. Richter, Y. Zhou, and D. Brunner (2011). "An Improved Tropospheric NO2 Column Retrieval Algorithm for the Ozone Monitoring Instrument". In: *Atmospheric Measurement Techniques* 4, pp. 1905–1928.

Bollineni-Balabay, O., J. van den Brakel, and F. Palm (2017). "State Space Time Series Modelling of the Dutch Labour Force Survey: Model Selection and Mean Squared Errors Estimation". In: *Survey Methodology* 43.1, pp. 41–67.

Bowman, K. O. and L. R. Shenton (1975). "Omnibus Test Contours for Departures from Normality Based on $\sqrt{b_1}$ and b_2". In: *Biometrika* 62.2, pp. 243–250.

Buchdahl, J. (2002). *Climate Change*. Fact Sheet Series for Key Stage 4 and A-Level. Atmosphere, Climate and Environment Information Programme, Manchester Metropolitan University, pp. 1–171.

Castle, J. L. and D. F. Hendry (2020). "Climate Econometrics: An Overview". In: *Foundations and Trends in Econometrics* 10.3-4, pp. 145–322.

Catania, L. and A. G. Billé (2017). "Dynamic Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances". In: *Journal of Applied Econometrics* 32.6, pp. 1178–1196.

Chen, R. and J. S. Liu (2000). "Mixture Kalman Filters". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.3, pp. 493–508.

Chen, R., J. S. Liu, and T. Logvinenko (2001). "A Theoretical Framework for Sequential Importance Sampling with Resampling". In: *Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science.* Springer, New York. Chap. 11, pp. 225–246.

Choi, H. and H. R. Varian (2009). “Predicting Initial Claims for Unemployment Benefits”. Working Paper, Available at SSRN: https://ssrn.com/abstract=1659307.

— (2012). “Predicting the Present with Google Trends”. In: *Economic Record* 88.SUPPL.1, pp. 2–9.

Cole, M. A., R. J. Elliott, T. Okubo, and Y. Zhou (2013). “The Carbon Dioxide Emissions of Firms: A Spatial Analysis”. In: *Journal of Environmental Economics and Management* 65.2, pp. 290–309.

Creal, D. (2012). “A Survey of Sequential Monte Carlo Methods for Economics and Finance”. In: *Econometric Reviews* 31.3, pp. 245–296.

Creal, D., S. J. Koopman, and A. Lucas (2011). “A Dynamic Multivariate Heavy-Tailed Model for Time-Varying Volatilities and Correlations”. In: *Journal of Business & Economic Statistics* 29.4, pp. 552–563.

— (2013). “Generalized Autoregressive Score Models with Applications”. In: *Journal of Applied Econometrics* 28.5, pp. 777–795.

Curier, R. L., R. Kranenburg, A. J. S. Segers, R. M. A. Timmermans, and M. Schaap (2014). “Synergistic Use of OMI NO2 Tropospheric Columns and LOTOS-EUROS to Evaluate the NOx Emission Trends Across Europe”. In: *Remote Sensing of Environment* 149, pp. 58–69.

D’Amuri, F. and J. Marcucci (2017). “The Predictive Power of Google Searches in Forecasting US Unemployment”. In: *International Journal of Forecasting* 33.4, pp. 801–816.

Dallmann, T. R. and R. A. Harley (2010). “Evaluation of Mobile Source Emission Trends in the United States”. In: *Journal of Geophysical Research Atmospheres* 115, pp. 1–12.

De Jong, P. (1989). “Smoothing and interpolation with the state-space model”. In: *Journal of the American Statistical Association* 84.408, pp. 1085–1088.

Deb, S. and R. S. Tsay (2019). “Spatio-temporal Models with Space-time Interaction and their Applications to air Pollution Data”. In: *Statistica Sinica* 29, pp. 1181–1207.

Delle Monache, D., I. Petrella, and F. Venditti (2016). “Adaptive State Space Models with Applications to the Business Cycle and Financial Stress”. CEPR Discussion Paper DP11599.

Doz, C., D. Giannone, and L. Reichlin (2011). “A Two-step Estimator for Large Approximate Dynamic Factor Models Based on Kalman filtering”. In: *Journal of Econometrics* 164.1, pp. 188–205.

Duncan, B. N., A. I. Prados, L. N. Lamsal, Y. Liu, D. G. Streets, P. Gupta, E. Hilsenrath, R. A. Kahn, J. E. Nielsen, A. J. Beyersdorf, S. P. Burton, A. M. Fiore, J. Fishman, D. K. Henze, C. A. Hostetler, N. A. Krotkov, P. Lee, M. Lin, S. Pawson, G. Pfister, K. E. Pickering, R. B. Pierce, Y. Yoshida, and L. D. Ziemba (2014). “Satellite Data of Atmospheric Pollution for U.S. Air Quality Applications: Examples of Applications, Summary of Data End-user Resources, Answers to FAQs, and Common Mistakes to Avoid”. In: *Atmospheric Environment* 94, pp. 647–662.

Durbin, J. and S. J. Koopman (2012). *Time Series Analysis by State Space Methods*. Second. Oxford Statistical Science Series. OUP Oxford.

— (1997). “Monte Carlo Maximum Likelihood Estimation for Non-Gaussian State Space Models”. In: *Biometrika* 84.3, pp. 669–684.

— (2002). “A Simple and Efficient Simulation Smoother for State Space Time Series Analysis”. In: *Biometrika* 89.3, pp. 603–615.

Elliott, G., T. J. Rothenberg, and J. H. Stock (1996). “Efficient Tests for an Autoregressive Unit Root”. In: *Econometrica* 64.4, pp. 813–836.

European Environment Agency (2018). *Nitrogen Oxides (NOx) Emissions*. Indicator Assessment.

Finlayson-Pitts, B. J. and J. N. Pitts Jr. (1993). “Atmospheric Chemistry of Tropospheric Ozone Formation: Scientific and Regulatory Implications”. In: *Air & Waste* 43.8, pp. 1091–1100.

Fowlie, M., E. Rubin, and R. Walker (2019). “Bringing Satellite-Based Air Quality Estimates Down to Earth”. In: *AEA Papers and Proceedings* 109, pp. 283–88.

Gagliardini, P., E. Ghysels, and M. Rubin (2017). “Indirect Inference Estimation of Mixed Frequency Stochastic Volatility State Space Models Using MIDAS Regressions and ARCH Models”. In: *Journal of Financial Econometrics* 15.4, pp. 509–560.

Gallant, A. R. and G. Tauchen (1998). “Reprojecting Partially Observed Systems with Application to Interest Rate Diffusions”. In: *Journal of the American Statistical Association* 93.441, pp. 10–24.

Giannone, D., L. Reichlin, and D. Small (2008). “Nowcasting: The Real-time Informational Content of Macroeconomic Data”. In: *Journal of Monetary Economics* 55.4, pp. 665–676.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). “A Novel Approach to Nonlinear and Non-Gaussian Bayesian State Estimation”. In: *IEE Proceedings. Part F: Radar and Sonar Navigation* 140.2, pp. 107–113.

Gourieroux, C. and A. Monfort (1993). “Pseudo-likelihood Methods”. In: *Handbook of Statistics 11*. Ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod. Elsevier Science Publishers B.V. Chap. 12, pp. 335–362.

Gourieroux, C., A. Monfort, and E. Renault (1993). “Indirect Inference”. In: *Journal of Applied Econometrics* 8, pp. 85–118.

Gourieroux, C., A. Monfort, and A. Trognon (1984). “Pseudo Maximum Likelihood Methods: Theory”. In: *Econometrica* 52.3, pp. 681–700.

Grange, S. K. (2014). *Technical Note: Averaging Wind Speeds and Directions*. Tech. rep., p. 12.

Guan, Y., M. C. Johnson, M. Katzfuss, E. Mannshardt, K. P. Messier, B. J. Reich, and J. J. Song (2019). “Fine-Scale Spatiotemporal Air Pollution Analysis Using Mobile Monitors on Google Street View Vehicles”. In: *Journal of the American Statistical Association*, pp. 1–14.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hansen, B. E. (2019). *Econometrics*. Jan 2019. University of Wisconsin, Department of Economics.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series*. Economic Series Monograph. Cambridge University Press.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Harvey, A. C. and C.-H. Chung (2000). “Estimating the Underlying Change in Unemployment in the UK”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163.3, pp. 303–309.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, US: Springer New York Inc.

Hastie, T. and H. Zou (2005). “Regularization and Variable Selection via the Elastic Net”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.

Hillebrand, E., S. Johansen, and T. Schmith (2020). “Data Revisions and the Statistical Relation of Global Mean Sea Level and Surface Temperature”. In: *Econometrics* 8.4, pp. 1–19.

Hindrayanto, I., S. J. Koopman, and J. de Winter (2016). “Forecasting and Nowcasting Economic Growth in the Euro Area Using Factor Models”. In: *International Journal of Forecasting* 32.4, pp. 1284–1305.

Hinrichsen, R. and E. E. Holmes (2009). “Using Multivariate State-space Models to Study Spatial Structure and Dynamics”. In: *Spatial Ecology*. Ed. by R. S. Cantrell, C. Cosner, and S. Ruan. CRC/Chapman Hall. Chap. 1, pp. 1–20.

Jungbacker, B. and S. J. Koopman (2007). “Monte Carlo Estimation for Nonlinear Non-Gaussian State Space Models”. In: *Biometrika* 94.4, pp. 827–839.

Kelejian, H. H. and I. R. Prucha (1998). “A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances”. In: *Journal of Real Estate Finance and Economics* 17.1, pp. 99–121.

Kitagawa, G. (1996). “Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models”. In: *Journal of Computational and Graphical Statistics* 5.1, pp. 1–25.

Koop, G. and D. Korobilis (2013). “Large time-varying parameter VARs”. In: *Journal of Econometrics* 177.2, pp. 185–198.

Koopman, S. J., K. M. Lee, and S. Y. Wong (2006). “Trend-Cycle Decomposition Models with Smooth-Transition Parameters: Evidence from U.S. Economic Time Series”. In: *Nonlinear Time Series Analysis of Business Cycles*. Contributions to Economic Analysis. Emerald Group Publishing Limited. Chap. 8, pp. 199–219.

Koopman, S. J., R. Lit, and T. M. Nguyen (2018). “Modified Efficient Importance Sampling for Partially Non-Gaussian State Space Models”. In: *Statistica Neerlandica* 73.1, pp. 44–62.

Koopman, S. J., A. Lucas, and M. Scharth (2015). "Numerically Accelerated Importance Sampling for Nonlinear Non-Gaussian State-Space Models". In: *Journal of Business and Economic Statistics* 33.1, pp. 114–127.

Krotkov, N. A., L. N. Lamsal, E. A. Celarier, W. H. Swartz, S. V. Marchenko, E. J. Bucsela, K. L. Chan, M. Wenig, and M. Zara (2017). "The version 3 OMI NO2 Standard Product". In: *Atmospheric Measurement Techniques Discussions* 10, pp. 3133–3149.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). "Supplementary Materials for The Parable of Google Flu: Traps in Big Data Analysis". In: *Science* 343.March, pp. 1203–1206.

Lee, L.-F. (2004). "Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models". In: *Econometrica* 72.6, pp. 1899–1925.

LeSage, J. and R. K. Pace (2009). *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press.

Li, M., S. J. Koopman, R. Lit, and D. Petrova (2020). "Long-term Forecasting of El Niño Events via Dynamic Factor Simulations". In: *Journal of Econometrics* 214.1, pp. 46–66.

Li, T., M. Bolić, and P. M. Djurć (2015). "Resampling Methods for Particle Filtering: Classification, Implementation, and Strategies". In: *IEEE Signal Processing Magazine* 32.3, pp. 70–86.

Lin, Y. and H. Reuvers (2020a). "Cointegrating Polynomial Regressions with Power Law Trends: A New Angle on the Environmental Kuznets Curve". Working paper, https://arxiv.org/abs/2009.02262.

— (2020b). "Efficient Estimation by Fully Modified GLS with an Application to the Environmental Kuznets Curve." Working paper, https://arxiv.org/abs/1908.02552.

Liu, J. S. and R. Chen (1998). "Sequential Monte Carlo Methods for Dynamic Systems". In: *Journal of the American Statistical Association* 93.443, pp. 1032–1044.

Ljung, G. M. and G. E. P. Box (1978). "On a Measure of Lack of Fit in Time Series Models". In: *Biometrika* 65.2, pp. 297–303.

Lopez-Aparicio, S., H. Grythe, R. J. Thorne, and M. Vogt (2020). "Costs and Benefits of Implementing an Environmental Speed Limit in a Nordic City". In: *Science of the Total Environment* 720, p. 137577.

Maas, B. (2019). *Short-term Forecasting of the US Unemployment Rate*. MPRA Paper 94066. University Library of Munich, Germany.

Maddison, D. (2006). "Environmental Kuznets Curves: A Spatial Econometric Approach". In: *Journal of Environmental Economics and Management* 51.2, pp. 218–230.

Merk, M. S. and P. Otto (2020). "Estimation of Anisotropic, Time-Varying Spatial Spillovers of Fine Particulate Matter Due to Wind Direction". In: *Geographical Analysis* 52.2, pp. 254–277.

Monfardini, C. (1998). "Estimating Stochastic Volatility Models through Indirect Inference". In: *The Econometrics Journal* 1.1, pp. C113–C128.

Montamat, G. and J. H. Stock (2020). "Quasi-experimental Estimates of the Transient Climate Response Using Observational Data". In: *Climatic Change* 160, pp. 361–371.

Moon, H. R. and B. Perron (2012). "Beyond Panel Unit Root Tests: Using Multiple Testing to Determine the Nonstationarity Properties of Individual Series in a Panel". In: *Journal of Econometrics* 169.1, pp. 29–33.

Naccarato, A., S. Falorosi, S. Loriga, and A. Pierini (2018). "Combining Official and Google Trends Data to Forecast the Italian Youth Unemployment Rate". In: *Technological Forecasting and Social Change* 130, pp. 114–122.

Nordhaus, W. (2013). *The Climate Casino: Risk, Uncertainty, and Economics for a Warming World*. Yale University Press, p. 392.

Ogen, Y. (2020). "Assessing Nitrogen Dioxide (NO2) Levels as a Contributing Factor to Coronavirus (COVID-19) Fatality". In: *Science of the Total Environment* 726, pp. 1–5.

Ozturk, I. (2015). "Measuring the Impact of Energy Consumption and Air Quality Indicators on Climate Change: Evidence from the Panel of UNFCC Classified Countries". In: *Environmental Science and Pollution Research* 22, pp. 15459–15468.

Pfeffermann, D. (1991). "Estimation and Seasonal Adjustment of Population Means Using Data from Repeated Surveys". In: *Journal of Business and Economic Statistics* 9.2, pp. 163–175.

Pfeffermann, D., M. Feder, and D. Signorelli (1998). "Estimation of Autocorrelations of Survey Errors with Application to Trend Estimation in Small Areas". In: *Journal of Business & Economic Statistics* 16.3, pp. 339–348.

Poirier, D. J. (1973). “Piecewise Regression Using Cubic Splines”. In: *Journal of the American Statistical Association* 68.343, pp. 515–524.

Proietti, T. and E. Hillebrand (2017). “Seasonal Changes in Central England Temperatures”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.3, pp. 769–791.

Rao, J. N. K. and I. Molina (2015). *Small Area Estimation*. 2nd ed. Wiley Series in Survey Methodology. John Wiley & Sons, Inc., p. 480.

Russell, A. R., L. C. Valin, and R. C. Cohen (2012). “Trends in OMI NO2 Observations over the United States: Effects of Emission Control Technology and the Economic Recession”. In: *Atmospheric Chemistry and Physics* 12, pp. 12197–12209.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. New York, NY, US: Springer-Verlag Publishing.

Schiavoni, C., S. J. Koopman, F. Palm, S. Smeekes, and J. van den Brakel (2021a). “Time-varying State Correlations in State Space Models and Their Estimation via Indirect Inference.” Tinbergen Institute Discussion Paper 2021-020/III.

Schiavoni, C., F. Palm, S. Smeekes, and J. van den Brakel (2021b). “A Dynamic Factor Model Approach to Incorporate Big Data in State Space Models for Official Statistics”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 184.1, pp. 324–353.

Schwarz, G. (1978). “Estimating the Dimension of a Model”. In: *Annals of Statistics* 6.2, pp. 461–464.

Shapiro, S. S. and M. B. Wilk (1965). “An Analysis of Variance Test for Normality (Complete Samples)”. In: *Biometrika* 52.3-4, pp. 591–611.

Shephard, N. and M. Pitt (1997). “Likelihood Analysis of Non-Gaussian Measurement Time Series”. In: *Biometrika* 84.3, pp. 653–667.

Shephard, N. G. and A. C. Harvey (1990). “On the Probability of Estimating a Deterministic Component in the Local Level Model”. In: *Journal of Time Series Analysis* 11.4, pp. 339–347.

Smith, P. L. (2008). “Splines As a Useful and Convenient Statistical Tool”. In: *The American Statistician* 33.2, pp. 57–62.

Stephens-Davidowitz, S. and H. Varian (2015). “A Hands-on Guide to Google Data”. Working Paper, https://people.ischool.berkeley.edu/ hal/Papers/2015/primer.pdf.

Stock, J. H. and M. W. Watson (2007). “Why Has U.S. Inflation Become Harder to Forecast?” In: *Journal of Money, Credit and Banking* 39.1, pp. 3–33.

— (1998). “Median Unbiased Estimation of Coefficient Variance in a Time-varying Parameter Model”. In: *Journal of the American Statistical Association* 93.441, pp. 349–358.

Suhoy, T. (2009). *Query Indices and a 2008 Downturn: Israeli Data*. Discussion Paper Series No. 2009.06. Bank of Israel, pp. 1–33.

TNO (2019). *Emissies en Depositie van Stikstof in Nenderland*. Factsheet. The Netherlands Organisation for applied scientific research.

Van den Brakel, J. and S. Krieg (2009). “Estimation of the Monthly Unemployment Rate Through Structural Time Series Modelling in a Rotating Panel Design”. In: *Survey Methodology* 35.2, pp. 177–190.

Van den Brakel, J. A. and S. Krieg (2016). “Small Area Estimation with State Space Common Factor Models for Rotating Panels”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179.3, pp. 763–791.

— (2015). “Dealing with Small Sample Sizes, Rotation Group Bias and Discontinuities in a Rotating Panel Design”. In: *Survey Methodology* 41.2, pp. 267–296.

Van der A, R. J., D. H. Peters, H. Eskes, K. F. Boersma, M. van Roozendael, I. De Smedt, and H. M. Kelder (2006). “Detection of the Trend and Seasonal Variation in Tropospheric NO2 over China”. In: *Journal of Geophysical Research Atmospheres* 111.12, pp. 1–10.

Wagner, M. (2015). “The Environmental Kuznets Curve, Cointegration and Nonlinearity”. In: *Journal of Applied Econometrics* 967, pp. 948–967.

Wilks, S. S. (1938). “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses”. In: *The Annals of Mathematical Statistics* 9.1, pp. 60–62.

Yamartino, R. J. (1984). “A Comparison of Several “Single-Pass” Estimators of the Standard Deviation of Wind Direction”. In: *Journal of Climate and Applied Meteorology* 23.9, pp. 1362–1366.

Yu, J., R. de Jong, and L.-f. Lee (2008). “Quasi-maximum Likelihood Estimators for Spatial Dynamic Panel Data with Fixed Effects when both n and T are Large”. In: *Journal of Econometrics* 146.1, pp. 118–134.

Impact paragraph

Chapters 2 and 3, and to a lesser extent Chapter 4, focus on the use of state space models to provide more timely, accurate and realistic estimates of Dutch unemployment. This is done by either using Big Data (Google Trends) and/or registry-based data (claimant counts), on top of survey-based data, or by employing the latter two types of data while accounting for their time-varying relationship. Clearly this research work can be beneficial for national statistical offices, who are in charge of publishing this kind of official statistics, and therefore interested in doing so in a precise and timely manner. Additionally, not all countries have data about claimant counts, whereas Google Trends are freely available virtually all over the world and therefore represent a very accessible source of information.

The methods employed in the above-mentioned chapters can also be used to estimate other macroeconomic variables, than unemployment, such as the gross domestic product of a country. Achieving accurate and timely estimates of these variables is important in order to have a better understanding of the current state of the economy, especially in times of economic distress. Moreover, modelling parameters as time-varying, which is relatively easily done with state space models, allows us to discover how the same variables can be affected by important events, such as the burst of a pandemic or economic and financial crises. All this knowledge can be used for policy making by, for instance, central banks.

The econometric methods discussed in Chapter 4 can, potentially, be used to model the same time-changing relationship discussed above, but the chapter itself is purely methodological. As already mentioned several times throughout the thesis, the main purpose of Chapter 4 is to illustrate the research process of trying several, unsuccessful, approaches to answer a research question, before finding the right one. Generally research papers only report the performance of the successful methods and therefore tend to hide all the work that has been done behind them. Sharing it can instead be helpful to provide a more realistic picture of what research entails and hopefully be relieving for researchers undergoing the same type of struggles. Moreover, it can prevent econometricians from taking the same wrong paths, or allow them to start from where we stopped in case they find a solution to our issues.

Chapter 5 illustrates how to use state space methods for modelling climate variables, specifically air pollutants. Climate change is obviously one of the biggest problems and challenges of our times and it is (fortunately) increasingly catching the attention of scientists from all kind of fields, including econometrics. Climate econometric models are thus becoming a complement to the widely used integrated assessment models in economics, which aim at understanding the relationship between economic and environmental variables. They allow to predict how future climate scenarios can affect the economy, or how future economic scenarios can affect the climate. Chapter 5 therefore brings a small contribution to this field by building an econometric model that can be used to evaluate the effect that hypothetical reductions in traffic intensity (and therefore a type of economic activity) have on pollution. This analysis can be relevant for makers of pollution-reduction policies.

About the author

Caterina Schiavoni was born on November 28, 1992 in Rome (Italy). She obtained a BSc in Statistics in July 2015 from Sapienza - Università di Roma (Italy), and a MSc in Econometrics in August 2016 from Maastricht University (The Netherlands). During her undergraduate studies she spent a semester at Université Paris 1 Panthéon-Sorbonne (France). After her studies she did a six-months internship in data analysis at UNHCR - The UN Refugee Agency, in Copenhagen (Denmark).

In April 2017 she joined the Department of Quantitative Economics at Maastricht University and the Methodology Department at Statistics Netherlands, as a PhD student under the supervision of Prof. Dr. Jan van den Brakel, Dr. Stephan Smeekes and Prof. Dr. Franz Palm. The findings of her research are documented in this doctoral dissertation and have been presented at several seminars and international conferences.

In spring 2020 she (virtually) visited the Center for Spatial Data Science of the University of Chicago (Illinois, USA) and as of April 2021 she is a researcher at Fondazione Eni Enrico Mattei, in Milan (Italy), working on the sustainability of firms and cities.