On mixed-measurement methods for producing economic statistics

Labonne, Paul

Awarding institution:
King's College London

On mixed-measurement methods for producing economic statistics

By Paul Labonne

Supervised by Prof. Weale

January 2022

A thesis presented for the degree of Doctor of Philosophy in Economics

King's College London

Abstract

This thesis explores new methods for producing economic statistics and their associated uncertainty. A particular focus is placed on techniques for studying conjointly heterogeneous economic series sampled on different frequencies and released asynchronously. Chapters 2 to 4 show how the British survey used to collect data on business turnover, a proxy for short-term GDP, may be replaced with value-added tax returns, which also contain information on business turnover. These chapters tackle a number of statistical issues related to temporal disaggregation, cleaning and forecasting using filtering techniques, specifically state space and score driven methods. There are three important takeaways from this work. First, an approximation to nonlinear temporal aggregation constraints which facilitates modelling and dissemination is illustrated. Second, it shows that a nonlinear approach for cleaning works better than the more-classical outlier detection strategy. Third, the resulting output estimates based on VAT returns exhibit a different profile than the survey currently in use. Tackling a different problem, Chapter 5 presents a new means for updating the uncertainty attached to US GDP nowcasts in a timely way by using prediction errors already observed in series related to economic growth. It shows how dependencies in the dispersion and asymmetry of prediction errors, which typically increase in the onsets of economic recessions, may be modelled across time series when these are not observed on the same frequency. The results show that this approach is particularly useful to capture the forecasting uncertainty observed during the onset of the coronavirus pandemic. While most of the thesis is concerned with economic growth statistics, Chapter 6 concludes with a study on migration statistics. Google queries on national newspapers are exploited to produce a timely indicator of migration flows using a novel quasi score driven model relying on a copula.

Acknowledgment

I am grateful to my supervisor Weale for our many discussions during the four years leading to this thesis. I thank ESCoE for the PhD scholarship as well as ONS for their collaboration on the VAT data. Finally, thank you Josie for your patience.

A paper based on the third chapter and co-authored with Weale has been published in the Journal of the Royal Statistical Society (Series A).

Contents

Chapter 1

Introduction

Economic statistics are used by policy makers, firms, households and citizens in general to make decisions and understand our society. For instance, growth domestic product (GDP), which is a measure of the domestic revenue generated by the economy and the leading measure of economic growth, is used by policy makers for monitoring welfare and controlling inflation. But economic statistics are generally not directly observable; instead they are constructed using a wide range of data sources from which we can extract social and economic phenomena using statistical methods.

While economic statistics take many forms, this thesis studies time series, that is data showing the evolution of a given variable over time. The main contributions lie in the presentation and illustration of new approaches for modelling jointly heterogeneous time series sampled on different frequencies. The methods are illustrated using data related to economic growth and migration.

Chapters 2 to 4 discuss how administrative data in the UK may be exploited to replace partially the survey used for deriving monthly GDP. The Office for National Statistics (ONS) base their short-term GDP estimate on the output measure of GDP. For this they collect monthly data on business turnover, the main proxy for output, using the monthly business survey (MBS). But firms also provide figures of their turnover when submitting their VAT returns to the tax authorities. These returns are filled by most firms in the UK, implying that they might be a better data source than the MBS. The very large sample of the VAT data could yield a more precise and granular picture of output across the economy.

But extracting monthly output figures from the VAT-based turnover data requires to solve a number of statistical issues. First, while ONS publish a monthly GDP estimate, the VAT data are rolling quarterly aggregates. Deriving monthly figures from quarterly figures is possible using temporal disaggregation methods, which is the subject chapter 3. The standard approach for temporal disaggregation in National Statistics Institutes consists of using least-squares methods. However, they do not yield satisfactory results when applied to the VAT data. Indeed, a special feature of the VAT data is that they are subject to very large measurement errors and these cannot be accounted for when using least-squares techniques; the resulting estimates therefore do not make sense.

To openly account for the measurement errors affecting the VAT data, Chapter 3 explores the use of an unobserved components framework using state space techniques. A particularly convenient approximation for modelling data in logarithms is discussed and illustrated. This approximation yields a simpler and more efficient alternative to the nonlinear approach typically employed when using state space methods for temporal disaggregation. Interestingly, the resulting monthly output estimates derived from the VAT data show a different historical trend than the MBS.

The ONS currently publish monthly output with a lag of about two months, but at this time only a relatively small proportion of the VAT returns have been submitted. To produce a timely picture of output using the VAT data it is necessary to forecast late returns; this is the subject of Chapter 4. By using successive vintages of the VAT data over time and modelling systemic revisions across these vintages it is possible to derive timely output estimates. However, the noise in the early vintages is much larger than the noise affecting the last vintage (the vintage analysed in the chapter 3). The data exhibit some very large outlying observations which tend to be positive only, violating the assumption of normally-distributed errors on which the Kalman filter relies for estimating state space models.

To handle the non-Gaussian features of the early vintages, Chapter 4 presents a score driven approach for temporal disaggregation. Unlike the Kalman filter, score driven models can handle non-Gaussian and non-linear features of the data. The results show that a score driven model for cleaning yields lower revisions on aggregate than the standard approach of discarding outliers through t-tests.

Chapter 5 explores the use of score driven techniques in a different, more general context. The focus is on probabilistic forecasts with an application to US economic growth data. Quarterly GDP in the US is published with a lag of a month, a significant delay if these numbers are to be used by policy makers to make timely decisions. But other series related to economic growth are released more frequently and rapidly than GDP, such as unemployment figures. By modelling the relationship between these related series and GDP, it is possible to improve current-quarter GDP forecasts each time new data on related series are released. This is the purpose of nowcasting methods.

But while nowcasting techniques make use of series related to economic growth for improving GDP point forecasts, the signal on forecasting uncertainty these series carry is generally not exploited fully. Chapter 5 presents a new means for updating nowcasting uncertainty following the release of related data, with a particular attention to the asymmetric behaviour of uncertainty during recessions. Using a novel score driven model it shows how cross-sectional dependencies in scale and shape parameters may be modelled in series sampled on different frequencies and released asynchronously. Scale and shape parameters control the dispersion and asymmetry characterising forecasting uncertainty; capturing their variations is especially important in times of economic stress. Modelling dependencies in scale and shape parameters proves to be especially useful for capturing the uncertainty arising in the onset of the coronavirus pandemic in the US in a timely way.

Chapter 6 ends this thesis with the presentation of a new approach to compute rapid estimates of net migration based on national newspaper queries searched on Google. We continue reading online newspapers from our home country even when we are abroad, and Google offers a view into this behaviour. For exploiting Google Trends data this chapter sets out a novel bivariate score driven model relying on a copula for capturing the joint conditional distribution of the observations. The approach is illustrated using New Zealanders net flow in Australia. A comparison with official migration statistics shows that the proposed method is precise.

Chapter 2

VAT returns as an alternative to the Month Business Survey

The United Kingdom has traditionally relied on surveys to provide the data needed to compile the national accounts. Following an independent review of UK economic statistics Bean (2016), however, there has been increased interest in the use of administrative data sources. A particular focus has been the use of Value Added Tax (VAT) returns to provide data on business turnover which are currently collected by the Monthly Business Survey (MBS). These data form the core of the output estimate of GDP, which has been published monthly by ONS since July 2018. Chapter 3 and 4 explore the role that VAT returns can play in delivering monthly estimates of business turnover, for use in generating output estimates of GDP.

ONS currently base its output estimate of GDP largely on data gathered with a monthly business survey, the MBS. It collects data on business turnover and is a sample survey with five strata. The first four strata are identified by employment size, as recorded in the Inter-Departmental Business Register. These are defined as employment groupings of 0 to 9, 10 to 49, 50 to 99 and 100 and over. Strata one to three are sampled while stratum four is fully enumerated. The fifth stratum is comprised of firms with turnover in excess of £60 million even though their employment size would put them in bands one to three. This stratum is also fully enumerated.

Separately, all firms with a turnover greater than £85,000 must be registered for VAT, and they are all required to declare their turnover as well as any tax due. The number of firms registered for VAT is many times larger than the sample of the MBS, suggesting that these returns might be a better data source for business turnover. But their administrative nature also creates challenges with their use in the production of monthly output figures. Hand (2018) discusses in detail common problems when using administrative data for statistical purposes. In the case of the VAT returns, there are issues of temporal aggregation (or frequency of aggregation), timeliness and measurement errors.

2.1 Temporal aggregation

Firms can submit their VAT returns annually, quarterly or monthly. Furthermore, firms reporting quarterly can start doing so in any month, generating three possible quarterly reporting patterns, referred to as quarterly *staggers* and represented in table 2.1. There is also one monthly stagger and twelve annual staggers made of firms reporting monthly and annual totals respectively, but the quarterly data are weighted to represent all of those. Thus the VAT figures analysed are rolling quarterly turnover figures.

Table 2.1: Representation of the quarterly staggers; x = quarterly turnover total

	Month												
	J	F	M	A	M	J	J	A	S	O	N	D	J
Stagger 1			x			x			x			x	
Stagger 2	x			x			x			x			x
Stagger 3		x			x			x			x		

In order to make use of these data it is necessary to interpolate monthly figures, to produce *interpolands*, from these rolling aggregates; this is the subject of chapter 3.

2.2 Measurement errors

The turnover figures extracted from VAT returns are subject to large measurement errors. And even though ONS apply three automatic rules for cleaning VAT returns (see Stephens and Allcoat (2016)), they remain very noisy.

Some of this noise could arise from the difficulty in linking VAT registration units to the business register used by ONS. If a VAT return relates to turnover arising from a number of distinct business activities, the turnover needs to be attributed between those activities. An issue arises as well when more than one VAT returns relate to a single business unit. ONS attribute VAT turnover to units from the business register using employment size. But some firms have complicated business patterns and carry out activities in several industries, making the attribution of turnover difficult. This problem of apportionment is more prevalent in large firms.

Overall the sources of measurement errors affecting the VAT returns are not well identified and understood. And the extremely large size of the VAT sample means that it is not possible to clean individual returns manually. As a result the VAT data analysed contain substantial measurement errors which complicate their interpolation. In this context least-squares techniques are not a viable means of interpolation. Consequently, an unobserved components model is developed in chapter 3 to accommodate the noise embedded in the figures as well as their overlapping nature.

2.3 Timeliness

Firms are required to submit their returns two months after the end of the period they relate to, i.e. firms submitting aggregate figures for the period January-March should submit at the end of April. Most firms, however, submit their returns with a delay.

Table 2.2: Comparison of the VAT and MBS data

	MBS	**VAT**
Nature	Survey	Administrative
Coverage	All large businesses and a sample of small firms	Almost all firms with an annual turnover of £85,000
Frequency	Monthly	Rolling quarterly
Timeliness	About 6 weeks	Small share after 6 weeks with many late returns
Noise	Almost null after initial cleaning	Very large measurement errors

Consequently the VAT data accrue gradually. Figure 2.1 illustrates the timeliness of the data. Typically, it takes five months for the VAT returns to be nearly complete; hence, most of the data relating to month t are available only at $t+5$. But revisions continue after although they tend to be small after twelve months. Chapter 4 shows how early figures can be used to generate timely forecasts of VAT-based turnover.

The MBS survey, on the other hand, yields monthly output figures with a two-month lag. Since the MBS covers all largest businesses (size bands 4 and 5) and is more timely than the VAT data, it will remain the measure of turnover for these businesses. This MBS series for the largest businesses can thus be used to improve estimation of monthly VAT-based turnover. The VAT data analysed below consequently cover only size bands one to three (small and medium size businesses).

2.4 Coverage of the datasets

Both the VAT and MBS datasets show data for seventy-five industries at a 2-digit Standard Industry Classification (SIC) level. These industries represent approximately a quarter of gross value added in the UK economy.

MBS figures start in January 2011 and VAT figures in March 2011. But both series are constructed using data from January 2011 because VAT figures are three-month totals . Chapter 3, which explores the temporal disaggregation problem, uses data up

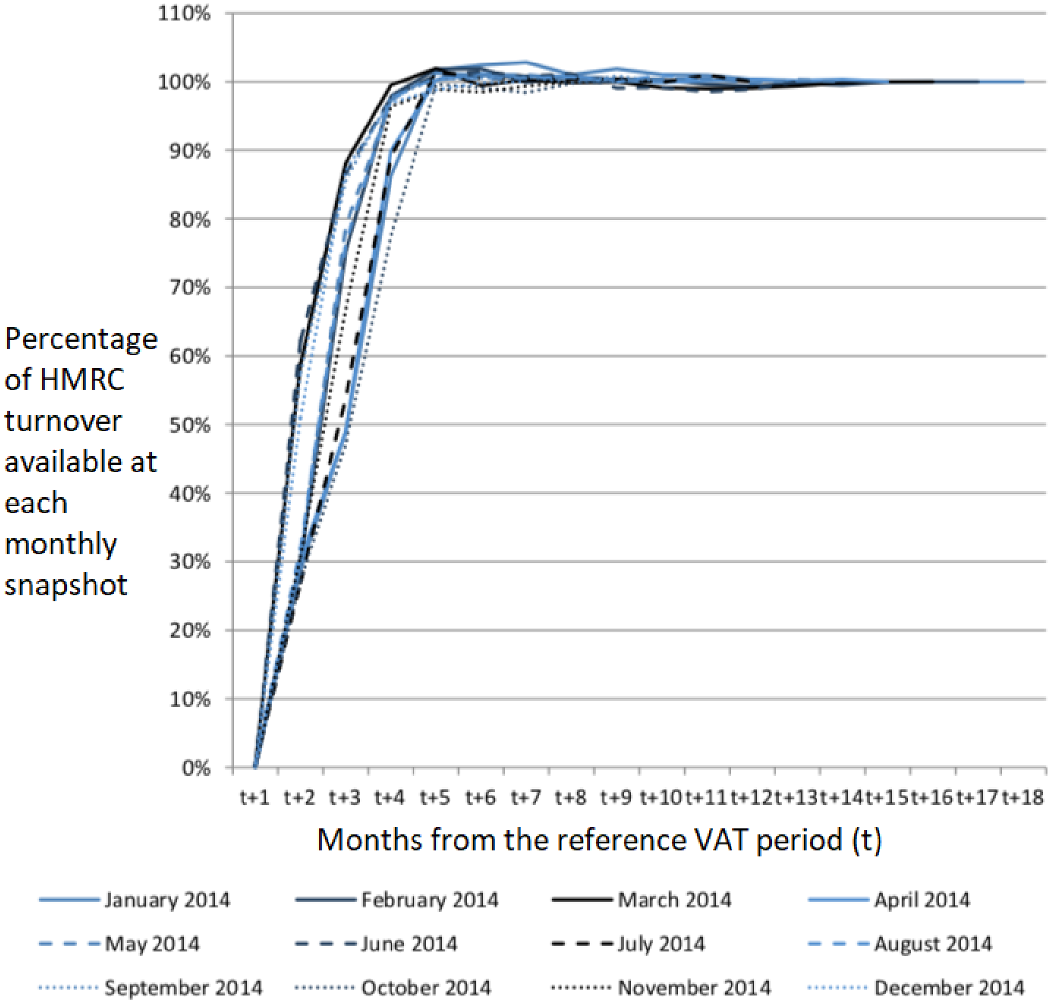


Figure 2.1: Timeliness of the VAT data from the reference period for 2014 for the whole economy. Source: Stephens and Allcoat (2016)

until July 2017. Chapter 4 uses data up until July 2020.

Chapter 3

Temporal disaggregation of noisy rolling quarterly data

3.1 Introduction

Most commonly used methods for temporal disaggregation, which are extensively exploited in national statistical institutes, are least-squares techniques. They rely on the constrained minimisation of a given residual (e.g the squared monthly deviation), where the observed aggregate totals serve as aggregation constraints. When potentially related series are available at the interpoland's frequency they can be used to improve estimation. These series are usually referred to as indicator, related or covariate series. Regression-based methods such as Chow and Lin (1971), Fernandez (1981) and Litterman (1983) make use of such series. If no related series are available at the desired frequency, the preferred method is Boot et al. (1967), which is equivalent to Fernandez's approach with a constant as indicator series. Mitchell et al. (2005) generalise least-squares techniques in a single framework and derive an approximation to use data in logarithms; below their approximation is applied to observations subject to seasonal movements. But least-squares techniques do not yield satisfactory results in the presence of measurement errors, and their application to the rolling quarterly VAT figures produces unrealistic estimates. When the observed data are noisy, the interpolands should not be constrained to the exact rolling quarterly figures, but instead to their underlying ‘clean’ signal. Consequently, this chapter adopts an unobserved components approach where the VAT observations are modelled as the sum of latent components, one of which is an observation error.

The MBS data derived from bands one to three are not used for estimating the monthly VAT-based output figures because the aim of this work is to investigate whether they can be replaced with the VAT data. However, they are used in two ways. First, they are used as a comparison metric for the interpolated figures because they should measure the same variable. Secondly, they are exploited to investigate how the model performs on ‘clean’ figures. For this they are artificially aggregated into three-month aggregates, which are subsequently disaggregated temporally with the model. Hence it is possible to compare the resulting interpolated series with the underlying true series.

After setting out the model and estimation strategy in the next section, the proposed method is illustrated using an example industry as a case study. This is useful to examine the goodness of fit of the model and show the effect of noise on estimation. Next a new monthly measure of turnover for small and medium size businesses is derived using the model on the seventy-five industries for which VAT returns are relevant. Aggregate estimates covering approximately a quarter of gross value added in the UK economy are shown and compared to the MBS figures.

3.2 An unobserved components approach

Unobserved components models can be estimated efficiently with state space techniques (see Harvey (1989) and Durbin and Koopman (2012)). For this the model is cast in state space form where the unobserved components are related to the observations through an observation equation. The overlapping and aggregated nature of the data can thus be accounted for through this observation equation. Moauro and Savio (2005) show how to make use of a covariate series in the Seemingly Unrelated Time Series (SUTSE) framework of Harvey and Koopman (1997). With this approach both the series subject to temporal disaggregation and the covariate series are modelled with unobserved components. In contrast to regression methods, SUTSE models rely on relatively weak assumptions regarding the form that the relationship between the interpoland and the covariate can take – it is assumed simply that both series are affected by common shocks. By adopting this approach both series can follow distinct trends. Furthermore the model can make use of covariates that are not seasonally adjusted by estimating seasonal effects for both the covariates and the interpolands simultaneously. The next section presents in detail the different components of the model, cast it in state space form, and describe the estimation strategy.

3.2.1 Modelling and estimation

This chapter develops a bivariate SUTSE model where the logs of the data are modelled as a sum of latent components. The VAT observations are modelled as the sum of the quarterly aggregation of a local linear trend, a quarterly seasonal effect, a calendar effect, a stagger bias, and, importantly, an observation error. The covariate observations (the MBS data for bands four and five), on the other hand, are not aggregated temporally and are free of measurement errors. They can be modelled as a sum of a local linear trend, a monthly seasonal effect and a calendar effect.

3.2.2 Observation equations

The exact observation function for the VAT figures is

$$y_{1,t} = \log(e^{x_{1,t}} + e^{x_{1,t-1}} + e^{x_{1,t-2}}) + \gamma_{1,t} + \beta_1 h^a_{1,t} + b^{(j)}_{1,t} + \epsilon^{(j)}_{1,t}, \tag{3.1}$$

for $t = 1, ..., N$, where $j = 1, 2, 3$, is a stagger identifier, $y_{1,t}$ is the log of the three-month VAT-based turnover aggregate, $x_{1,t}$ is the log of the seasonally adjusted interpoland, $\gamma_{1,t}$ is a rolling quarterly seasonal effect, β_1 is an Easter effect, $h^a_{1,t}$ an Easter dummy and $b^{(j)}_{1,t}$, captures the possible bias of stagger j. Reporting errors or other types of measurement errors arising from the processing of the data are captured with $\epsilon^{(j)}_{1,t}$, which are modelled as white noise with stagger-specific variance:

$$\epsilon^{(j)}_{1,t} \sim N(0, \sigma^2_{1,\epsilon j}). \tag{3.2}$$

The three-month VAT turnover aggregates are observed from the third period onward; hence the first two values of $y_{1,t}$ are missing.

The observation equation for the covariate is

$$y_{2,t} = x_{2,t} + \gamma_{2,t} + \beta_2 h^a_{2,t}, \tag{3.3}$$

for $t = 1, ..., N$, where $y_{2,t}$ is the log of the monthly covariate, $\gamma_{2,t}$ is a monthly seasonal effect, $h^a_{2,t}$ an Easter dummy and β_2 an Easter effect. This observation equation does not involve any time aggregation because the monthly values of the covariate are observed directly.

The temporal aggregation technique used here differs from the widely used method introduced by Harvey and Pierse (1984) (see Grassi et al. (2015) for a recent application) and does not rely on the augmentation of the state vector with a cumulator variable.

This approach suits the rolling nature of the data better and is also generalisable to non-overlapping figures. This overlapping aggregation approach is common in the now-casting literature (see for instance Aruoba et al. (2009)). To understand the benefit of using an overlapping aggregation function like (3.1) instead of a cumulator variable it is helpful to compare the two methods with a simple example.

Comparison with Harvey and Pierse and generalisation

The approach is illustrated by considering the problem of estimating an unobserved monthly vector

$$\boldsymbol{m} = (m_1, m_2, ..., m_t, ..., m_N)'$$

where N is the total number of months. The quarterly observations are stored in the vector

$$\boldsymbol{q} = (., ., q_3, ., ., q_6, ., ., q_9, ., ., ., q_N)'$$

where $q_t = m_t + m_{t-1} + m_{t-2}$ for $t = 3, 6, 9, ..., N$, and dots indicate missing values. The difference between the two methods stems in the way these missing values are modelled. For simplicity both $\boldsymbol{m}$ and $\boldsymbol{q}$ are in levels. Harvey and Pierse model the observations such that

$$q_t = \psi_t q_{t-1} + m_t, \quad \psi_t = \begin{cases} 0, & \text{if } t = 3(\tau - 1) + 1 \\ 1, & \text{otherwise,} \end{cases}$$

where $\tau = 1, 2, ..., N/3$. For instance, the first missing value would be modelled as m_1, the second as $m_1 + m_2$, the third as the m_4 and so on. To achieve this it is necessary to augment the state vector with a cumulator variable and the transition matrix must be time-varying for the cumulator to 'reset' at the beginning of each quarter.

Here the state vector is not augmented with a cumulator variable. Instead, it is the observation function, or observation matrix in the case of linear models, that serves as the cumulator. This amounts to modelling the observations as

$$q_t = \sum_{i=0}^{2} m_{t-i}.$$

Hence, while Harvey and Pierse (1984) model the missing values as ‘partial cumulants’, here they are modelled as rolling 3-month totals.

The overlapping aggregation approach offers several benefits such as the direct observability of the interpoland estimates and their standard errors. Augmenting the state vector with a cumulator variable requires a decumulation step post estimation, while a second augmentation is required to observe the interpolands’ standard errors. However, with the overlapping aggregation function the dimension of the state vector can sometimes become an issue, for instance when modelling daily data. This is because it is necessary to add to the state vector all the lags present in the aggregation constraint. If each lag is the sum of multiple model components, such as a trend and an irregular component, the dimension of the state vector can increase significantly; in this case using a cumulator variable might be preferable.

3.2.3 Linearisation strategy

The observation equation (3.1) is nonlinear whereas the Kalman filter, which is used to evaluate the model log likelihood for estimation, relies on linear models. Proietti and Moauro (2006) present a *sequential linear constrained* (SLC) method for estimating state space models with nonlinear aggregation constraints. This algorithm is explained further in Proietti (2006) in the general context of nonlinearly aggregated mixed models. When the model is Gaussian, which is the case here, this approach is equivalent to the linearisation by mode estimation as set out in Durbin and Koopman (2012). The SLC algorithm is an iterative method which consists of estimating a model with an approximate linear constraint, where the approximation is sequentially improved using the solution of the estimation as a new guess until convergence. The approximation error can thus be reduced to zero. This is a critical feature for national statistical institutes; they attach a lot of importance to respecting aggregation constraints exactly. Appendix C provides details on the derivation of the approximated linear model and the SLC algorithm.

An alternative approach could be to use the Extended Kalman filter. This method consists of running the Kalman filter on the nonlinear model, where at every step of the filter the observation equation is linearised with a first order Taylor approximation at the one step ahead prediction that is yielded by the previous step, with the initial value of the state vector as an approximation point for the first step. This approach has two significant disadvantages compared to the SLC method. First, it cannot be readily employed with statistical packages because few offer an Extended Kalman filter routine. Secondly, and more importantly, it would yield discrepancies in the aggregation constraints because the approximation errors cannot be reduced to a chosen tolerance value. These discrepancies can be significant if the month-on-month changes in the interpolands are large.

While the SLC algorithm is cumbersome to implement, the Extended Kalman filter is not precise enough. Mitchell et al. (2005) propose a third approach relying on a precise approximation of the temporal aggregation constraint. Using this approximation it is possible to retain a linear formulation of the model and thus the standard Kalman filter can be used for estimation. This approach is nearly as precise as the iterative method of Proietti and Moauro (2006) while being considerably simpler. The approximation is the following:

$$\sum_{i=0}^{2} \mathrm{h}(z_{t-i}) \approx 3\ \mathrm{h}\left(\frac{\sum_{i=0}^{2} z_{t-i}}{3}\right), \tag{3.4}$$

where $h(.)$ is a nonlinear function and z_t a relatively smooth variable comparable to standard macroeconomic data. Appendix A provides details on the derivation. Using approximation (3.4) with the exponential function it is possible to approximate precisely (3.1) with

$$y_{1,t} = \ln 3 + \frac{1}{3}x_{1,t} + \frac{1}{3}x_{1,t-1} + \frac{1}{3}x_{1,t-2} + \gamma_{1,t} + \beta_1 h^a_{1,t} + b^{(j)}_{1,t} + \epsilon^{(j)}_{1,t}. \tag{3.5}$$

This approximation is possible only if the seasonal movements are estimated as quarterly seasonal movement because the the approximation relies on the monthly movements being relatively small.

3.2.4 A bivariate local linear trend for the seasonally adjusted figures

The logs of the seasonally adjusted covariate and interpoland follow a bivariate local linear trend model:

$$\begin{aligned} x_t &= \mu_t + e_t, & e_t &\sim \mathrm{N}(0, \Sigma_e), \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim \mathrm{N}(0, \Sigma_\xi), \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim \mathrm{N}(0, \Sigma_\zeta), \end{aligned} \tag{3.6}$$

where $x_t = (x_{1,t}, x_{2,t})'$, $\mu_t = (\mu_{1,t}, \mu_{2,t})'$ is the vector of dynamic trends, $\nu_t = (\nu_{1,t}, \nu_{2,t})'$ is the vector of dynamic slopes and $e_t = (e_{1,t}, e_{2,t})'$ is the vector of irregular components. Unlike the measurement error, the irregular component carries economic meaning, although both are modelled as white noise. The irregular variation in the covariate are used to help separating the irregular component in the interpoland from the measurement error. The vectors $\xi_t = (\xi_{1,t}, \xi_{2,t})'$ and $\zeta_t = (\zeta_{1,t}, \zeta_{2,t})'$ refer to the disturbances of the trend and slope components respectively.

It is assumed that the disturbances are uncorrelated across time and across unobserved components, but there can be a contemporaneous correlation within each unobserved component. It is through these contemporaneous correlations that the covariate can be useful in estimating the interpoland. Specifically, the covariance matrix Σ_h, $h = \xi, \zeta, e$, is defined as

$$\Sigma_h = \begin{pmatrix} \sigma^2_{1,h} & \rho_h \sigma_{1,h} \sigma_{2,h} \\ \rho_h \sigma_{1,h} \sigma_{2,h} & \sigma^2_{2,h} \end{pmatrix},$$

with $\sigma^2_{1,h}$ the variance of the interpoland's h component and $\sigma^2_{2,h}$ the variance of the covariate's h component.

3.2.5 Seasonality and stagger bias

There are two important issues arising when estimating seasonality in noisy rolling quarterly figures. First, it is not possible to estimate monthly seasonal effects from quarterly observations because there is more than one set of monthly effects consistent with the estimated quarterly effects. The overlapping nature of the data does not help because, although twelve quarterly seasonal effects are estimated, two of them are linearly dependent on the other ten, leaving the problem indeterminate; appendix B provides more details on this point. But even though seasonality cannot be estimated at the disaggregate (monthly) level, it can be estimated at the aggregate (quarterly) level, which is enough to derive a seasonally adjusted interpoland.

Secondly, the large measurement errors in the aggregate data complicate estimation of the changes in the rolling quarterly seasonal effects because both measurement errors and seasonal disturbances are generally modelled as white noise. For this reason it is preferable to use a deterministic rolling quarterly seasonal model for the aggregate data such that

$$\gamma_{1,t} = -\gamma_{1,t-2} - \gamma_{1,t-5} - \gamma_{1,t-8}, \tag{3.7}$$

where $\gamma_{1,t}$ is a three-month seasonal effect.

The seasonality in the covariate, which unlike in the VAT data, is observed monthly and is not subject to measurement errors, can be captured with a stochastic trigonometric model:

$$\gamma_{2,t} = \sum_{j=1}^{6} \gamma_{2,j,t}, \qquad \begin{bmatrix} \gamma_{2,j,t+1} \\ \gamma^*_{2,j,t+1} \end{bmatrix} = \begin{bmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{bmatrix} \begin{bmatrix} \gamma_{2,j,t} \\ \gamma^*_{2,j,t} \end{bmatrix} + \begin{bmatrix} \omega_{2,j,t} \\ \omega^*_{2,j,t} \end{bmatrix}, \tag{3.8}$$

where $\lambda_j = 2\pi j/12$, for $j = 1, ..., 5$, and $\gamma_{2,6,t+1} = -\gamma_{2,6,t} + \omega_{2,6,t}$. The disturbances $\omega_{2,j,t}$ and $\omega^*_{2,j,t}$ are independently and normally distributed with zero means and variance $\sigma^2_{2,\omega}$ for $j = 1, ..., 5$, and $\sigma^2_{2,\omega}/2$ for $j = 6$; for a detailed exposition of the trigonometric model and a comparison with other seasonal models see Proietti (2000).

Changes to the rolling quarterly seasonal pattern in the VAT data could, to some extent, be borrowed from the covariate using the common seasonal model of Moauro and Savio (2005). But because of the typical slow evolution of seasonality and the limited sample size it is not helpful here. Nevertheless, the logarithmic specification implies that the seasonal effects captured are proportionate effects, which should change more slowly over time than additive effects, making the assumption of fixed seasonality in the aggregate figures reasonable.

Separately, the stagger biases are modelled explicitly using a dynamic specification taking the form of random walk:

$$b^{(j)}_{1,t+1} = b^{(j)}_{1,t} + \kappa^{(j)}_{1,t}, \qquad \kappa^{(j)}_{1,t} \sim N(0, \sigma^2_{1,\kappa j}), \quad j = 2, 3. \tag{3.9}$$

It is not possible to identify a bias in all of the three staggers, so the biases in the second and third staggers are defined with respect to the first stagger, whose bias is fixed to zero. The covariate is not subject to this bias.

3.2.6 Calendar variations

Calendar effects can be classified in two broad categories: moving festivals and trading days effects. The Easter period is the only significant moving festival in the UK. Easter can fall either in March or in April and can overlap both months. The common approach to estimate the Easter effect is due to Bell and Hillmer (1983) and consists of setting $E_t = \beta h_t$, with E_t the Easter effect at period t, and h_t the proportion of the total number of days in the Easter period (H_t) that falls in month t. For the Easter effects to add up to zero over a year, a representation similar to Harvey (2006) is adopted:

$$E_{i,t} = \beta_i(h_{i,t} - \sum_{t=1}^{s} h_{i,t}/s) = \beta_i h^a_{i,t}, \qquad i = 1, 2, \tag{3.10}$$

where s is the frequency, which is twelve here, and $\sum_{t=1}^{s} h_{2,t} = 1$ while $\sum_{t=1}^{s} h_{1,t} = 3$ (the three-month aggregates ending in April and May include both March and April; hence two months for which h_t is equal to one).

In the US, businesses can see their turnover fluctuate considerably round the seven days preceding Easter and it is usual to set $H_t = 7$. However, Russ and Tariq (2017) suggest using a specific Easter period for the UK. Their results show that it is better to account for the period Good Friday to Easter Monday, the entire bank holiday period, or from Monday before Easter Sunday to Friday following it. Here the former is chosen and H_t is set to four.

The trading days effects are usually insignificant with quarterly data because the varying number of trading days in a month partially averages out over three months (X-13ARIMA-SEATS reference manual, US Census Bureau).

3.2.7 State space representation and estimation

The observation equations (3.5) and (3.3), and the equations governing the unobserved components (3.2) and (3.6) - (3.10) can be cast together in state space form as

$$\begin{aligned} y_t &= Z_t \alpha_t, \\ \alpha_{t+1} &= T\alpha_t + R\eta_t, \qquad \eta_t \sim \mathrm{N}(0, Q), \\ \alpha_1 &\sim \mathrm{N}(a_1, P_1), \end{aligned} \tag{3.11}$$

where $y_t = (y_{1,t} - \log 3, y_{2,t})'$ and $\alpha_t = (\alpha'_{1,t}, \alpha'_{2,t})'$;
$\alpha_{1,t} = (\mu_{1,t}, \mu_{1,t-1}, \mu_{1,t-2}, \nu_{1,t}, \gamma_{1,t}, ..., \gamma_{1,t-8}, e_{1,t}, e_{1,t-1}, e_{1,t-2}, \beta_1, b^{(2)}_{1,t}, b^{(3)}_{1,t}, \epsilon^{(1)}_{1,t}, \epsilon^{(2)}_{1,t},$ $\epsilon^{(3)}_{1,t})'$, $\alpha_{2,t} = (\mu_{2,t}, \nu_{2,t}, \gamma_{2,1,t}, \gamma^*_{2,1,t}, \gamma_{2,2,t}, \gamma^*_{2,2,t}, ..., , \gamma_{2,5,t}, \gamma^*_{2,5,t}, \gamma_{2,6,t}, e_{2,t}, \beta_2)'$. Appendix D.1 shows the full representation of the state space form. Notice that the measurement errors are included in the state vector.

The unknown variance parameters are estimated via maximum likelihood, where the log likelihood function is evaluated with the prediction error decomposition using the Kalman filter. The Kalman filter is a recursion initialised with the mean vector a_1 and covariance matrix P_1 of the initial state vector which are unknown. The trends, slopes, seasonal effects, Easter effects and stagger biases are initialised with an exact diffuse initialisation, that is to say with arbitrary means and infinite variances, while the stationary states (the irregular component and the observations errors) are initialised with zero means and their unconditional variances.

While the Kalman filter yields optimal predictions of the state vector, which is suited to estimation and forecasting, the estimate of the state vector and its error variance given the entire sample are obtained with the Kalman smoother. The Kalman smoother is a backward recursion which makes use of the Kalman filter output when the parameters are set to their maximum likelihood estimates. The interpolands in levels are constructed from the smoothed estimates of the states as

$$\exp(\hat{x}_{1,t}) = \exp(\hat{\mu}_{1,t} + \hat{e}_{1,t}), \qquad t = 1, ..., N. \tag{3.12}$$

The smoothed estimates are used and not the filtered estimates because the latter typically lag the true series. While this would not be an issue if the objective was to forecast the VAT data, filtered estimates are not suitable for constructing an historical series.

The Kalman filter and smoother algorithms with exact diffuse initialisation of Durbin and Koopman (2012) and the associated log likelihood function are used. The model is estimated in R with the KFAS package (Helske (2017)) which contains the necessary functions for the filter, smoother and likelihood evaluation.

3.2.8 Detection and treatment of outlying observations

Extreme observation errors affecting the VAT data can have repercussions on the variance estimates and possibly affect the trend in the interpoland. Therefore, these large errors should be detected and treated, which is done through the analysis of the standardised state disturbances given by

$$\hat{\eta}_t^s = B_t^{\eta} \hat{\eta}_t, \qquad \text{with} \quad B_t^{\eta'} B_t^{\eta} = [\text{Var}(\hat{\eta}_t)]^{-1}, \tag{3.13}$$

where $\hat{\eta}_t$ is the vector of smoothed state disturbances and $\text{Var}(\hat{\eta}_t)$ its error variance matrix; both are outputs of the Kalman smoother.

Each standardised state disturbance can be tested for significance using a two-tailed t-test with the null hypothesis that the corresponding observation is not an outlier. It is common to choose a confidence interval of 95% which is associated with critical values of +/- 1.96. However, this is a multiple testing problem and on average $76/20 \approx 4$ outliers should be detected by chance. Replacing these observations with missing values might generate a substantial loss of information. Because of the large number of industries studied here, it is not possible to analyse each outlying observation individually and decide whether or not it should be excluded. Therefore, a more conservative confidence interval of 99.9% is adopted which is associated with critical values of ±3.3. The outlying observations detected are replaced with missing values. It is possible to account for this type of outlying observations with dummies, but such an approach is less parsimonious and yields similar results.

To accommodate the resulting missing values in the VAT series the algorithms for univariate treatment of multivariate series of Durbin and Koopman (2012) are used. With these, missing values are accommodated by changing the dimension of the observation vector which allows use of information in the covariate even in the months where the VAT figures are missing.

3.3 An application: Land transport and transport service via pipelines

The method for temporal disaggregation of the VAT figures is illustrated with the 'Land transport and transport services via pipelines' industry (SIC 493T495). Turnover of firms in size bands one to three accounts for approximately 45% of total turnover in this industry, and it has a weight of 1.5% in overall gross value added, making it one of the most important industries in our sample.

Figure 3.1 shows the industry's raw data. They include the VAT rolling quarterly data for bands one to three subject to temporal disaggregation, the MBS data for bands four and five that are used as covariates, and the MBS data for bands one to three that the VAT figures should replace.

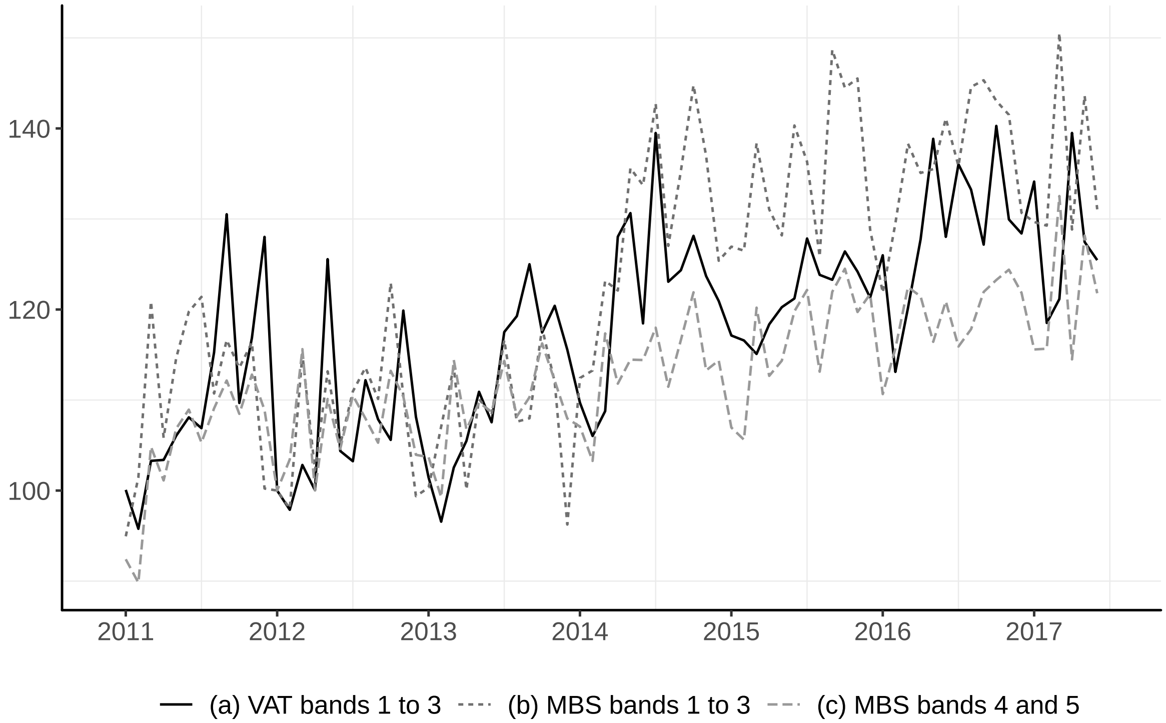


Figure 3.1: Raw data from industry 493T495, non-seasonally adjusted series, index January 2012 = 100

First the synthetic data constructed from the MBS data for bands one to three are used for estimation. The resulting interpolands can be compared to the true figures and it is thus possible to observed the efficiency of the proposed method when applied to clean data. Next the VAT data are analysed and the covariate is shown to be helpful. Finally the interpolated figures from the VAT data are compared with the MBS figures.

An important issue when estimating the VAT figures concerns the smoothness of the interpolands. This stems from the difficulty in separating the monthly changes in the seasonally adjusted figures from the measurement errors. Consequently, a measure of volatility in the interpolands is used as a performance indicator in addition to the log likelihood. Specifically, the root mean square log deviation in the seasonally adjusted monthly figures is calculated:

$$\text{RMSD} = \sqrt{\frac{1}{n-1}\sum_{t=2}^{n}(\hat{x}_{1,t} - \hat{x}_{1,t-1})^2}.$$

Obviously there is no particular value this statistic ‘ought’ to take, but it is helpful to be able to compare it with the comparable figure for the MBS, in order to indicate whether the interpolated series is more or less volatile than the MBS data.

3.3.1 Estimation with synthetic data

The synthetic figures are constructed by aggregating the MBS for bands one to three into three-month rolling aggregates. The aim of using a synthetic series is to compare the interpolated series with the underlying ‘true’ series. Specifically, the interpolation error is measured as the root mean square difference between the two series in logs. However, the interpolands are seasonally adjusted while the MBS data are raw figures. To remedy this issue the MBS figures are seasonally adjusted using the seasonal model applied to the VAT data, that is a deterministic seasonal model. Thus, it is possible to observe directly the efficiency of the interpolation procedure because both the MBS figures and the interpolands are subject to the same seasonal adjustment method.

Model (3.11) is estimated with the synthetic series and no covariate. A first round of estimation suggested a constant slope; therefore the disturbances in the slope are fixed to zero. The results of estimation are presented in the first row of table 3.1. Figure 3.2 shows the unobserved components’ smoothed estimates. The observation error variances of the second and third staggers are slightly positive, although the synthetic figures are, by construction, not subject to noise. But these observation errors are small and are likely to reflect changes in the seasonal pattern which cannot be captured with constant seasonal effects. Separately, as should be expected, the stagger bias is close to zero for both staggers. With a t-value of -1.05 the Easter effect is not statistically different from zero, and this is the case for all the subsequent estimation results.

Figure 3.3 compares the seasonally adjusted monthly interpolands in levels to the true series. The interpolands follow closely the underlying true figures, and this is

Table 3.1: Estimation results. Rows (1) refers to estimation with synthetic data, rows (2) to 'univariate' estimation (where the correlations are constrained to zero) and rows (3) to multivariate estimation.

	$\sigma_{\mu,1}$	$\sigma_{\nu,1}$	$\sigma_{e,1}$	$\sigma_{\epsilon,1}$	$\sigma_{\epsilon,2}$	$\sigma_{\epsilon,3}$	$\sigma_{b,2}$	$\sigma_{b,3}$	
(1)	0.44	0	0.878	0	0.002	0.017	0	0	
(2)	0.331	0	0.008	0.787	1.739	0.846	0.165	0.324	
(3)	0.352	0	0.06	0.9	1.727	0.825	0.142	0.274	
	$\sigma_{\mu,2}$	$\sigma_{\omega,2}$	$\sigma_{\nu,2}$	$\sigma_{e,2}$	ρ_{ξ}	ρ_{ζ}	ρ_{e}	Log lik	RMSD
(1)	-	-	-	-	-	-	-	228.048	0.04
(2)	0.054	0	0	0.627	0	-	0	188.792	0.008
(3)	0.065	0	0	0.608	0.679	-	-0.897	189.823	0.012

confirmed with an interpolation error of 0.013. With a root mean square log deviation of 0.04, the interpolated series is slightly smoother than the true series which exhibits a root mean square log deviation of 0.052. State space estimation typically generates a smoother series.

The shaded area represents the 90% confidence interval for the interpoland constructed as

$$\exp\left[\hat{x}_{1,t} \pm 1.65\sqrt{\mathrm{Var}(x_{1,t}|Y_n)}\right], \tag{3.14}$$

where $\mathrm{Var}(x_{1,t}|Y_n)$ is the error variance of the interpoland which is derived from the Kalman smoother recursion. This confidence interval reflects the filtering and smoothing uncertainty and thus is useful to convey partially the *epistemic uncertainty* about the modelling framework. This uncertainty comes from the lack of knowledge concerning the form of the model and its parameters. By contrast, analysing artificially aggregated figures, and thus being able to compare the interpolands with the underlying true figures, is useful to reduce the *ontological uncertainty* which concerns the model's accurate description of reality. Note that an additional source of ontological uncertainty arises when studying the actual VAT figures because they are subject to measurement errors whose statistical behaviour are largely unknown. For a detailed discussion about uncertainty and ways of communicating it see Spiegelhalter (2017).

The accuracy of the approximation to the temporal aggregation constraint can be evaluated by aggregating the unobserved components' smoothed estimates using the exact observation function (3.1) and comparing the resulting figures with the logs of the aggregated data; the difference between the two representing the approximation errors. Estimation with the synthetic series yields a root mean square approximation error of $3.7e^{-4}$, showing the very high precision of the approximation.

Overall, estimation results with the synthetic data show that the model is satisfactory for disaggregating temporally rolling quarterly turnover figures that are not seasonally adjusted. The VAT rolling quarterly series, which unlike the synthetic data exhibit measurement errors, are estimated in the next section.

3.3.2 Estimation with VAT data

Having demonstrated the goodness of fit of the estimation method with a synthetic series, it is now applied to the rolling quarterly VAT figures. The second row of table 3.1 shows the estimates resulting from estimating (3.11) with correlation coefficients between the VAT and covariate components fixed to zero. This is equivalent to estimating the VAT data with a univariate model. The third row of the table shows the results when the two series are allowed to be correlated. Thus it is possible to understand the implications of using a covariate when estimating the interpolands. In both cases the covariate's seasonal disturbances and the disturbances in the slopes were close to zero; hence deterministic slopes and seasonal effects are estimated. Figure 3.4 shows the smoothed estimates of the unobserved components for the VAT series arising from the unconstrained multivariate estimation.

The estimated correlation coefficient in the trends is positive while the irregular components are estimated to be negatively correlated. Importantly, using the covariate for estimating the interpoland leads to an improvement in the log likelihood and an

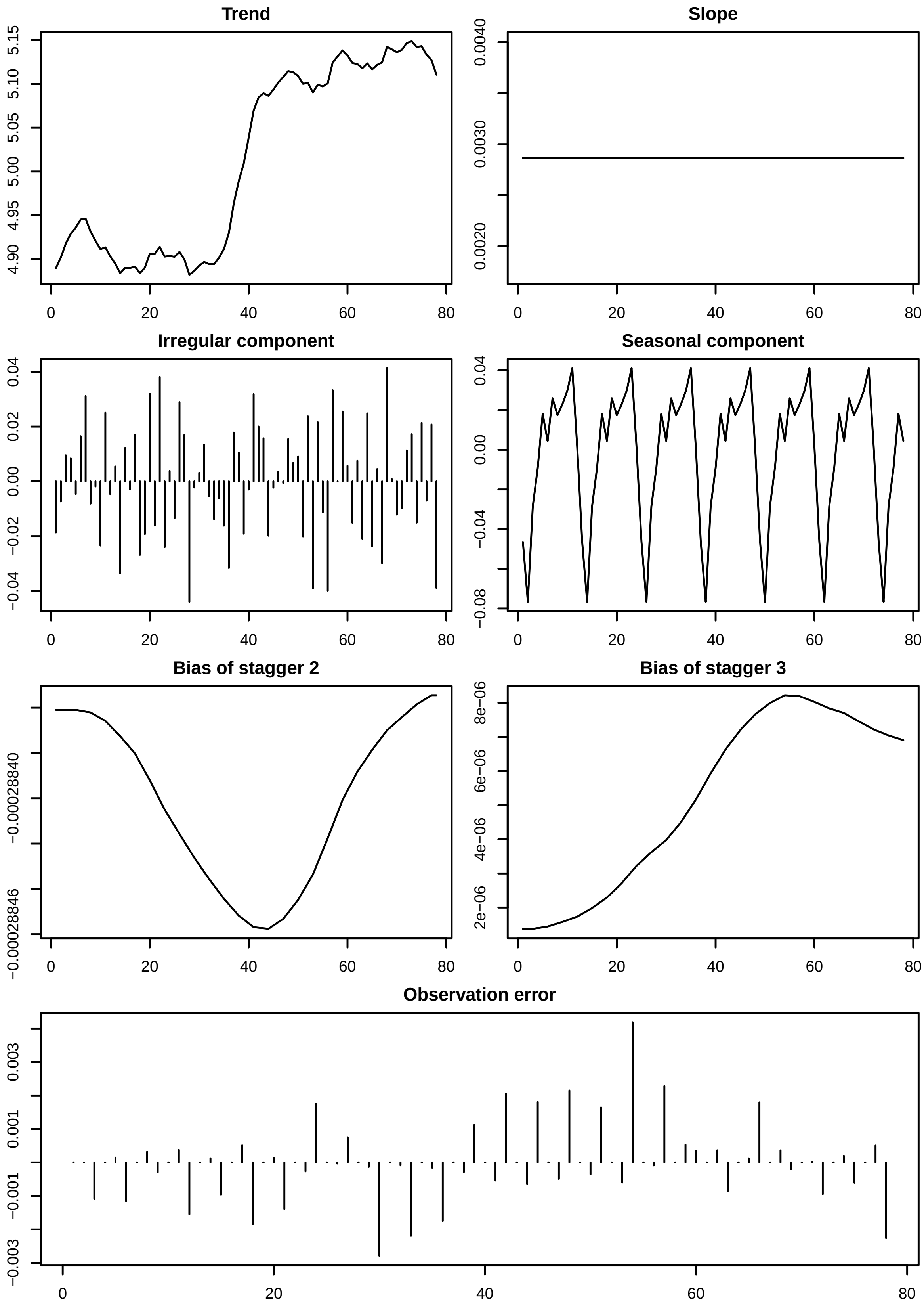


Figure 3.2: Smoothed unobserved components from estimation of the synthetic data constructed from MBS 1-3. Data from industry 493T495. Log units.

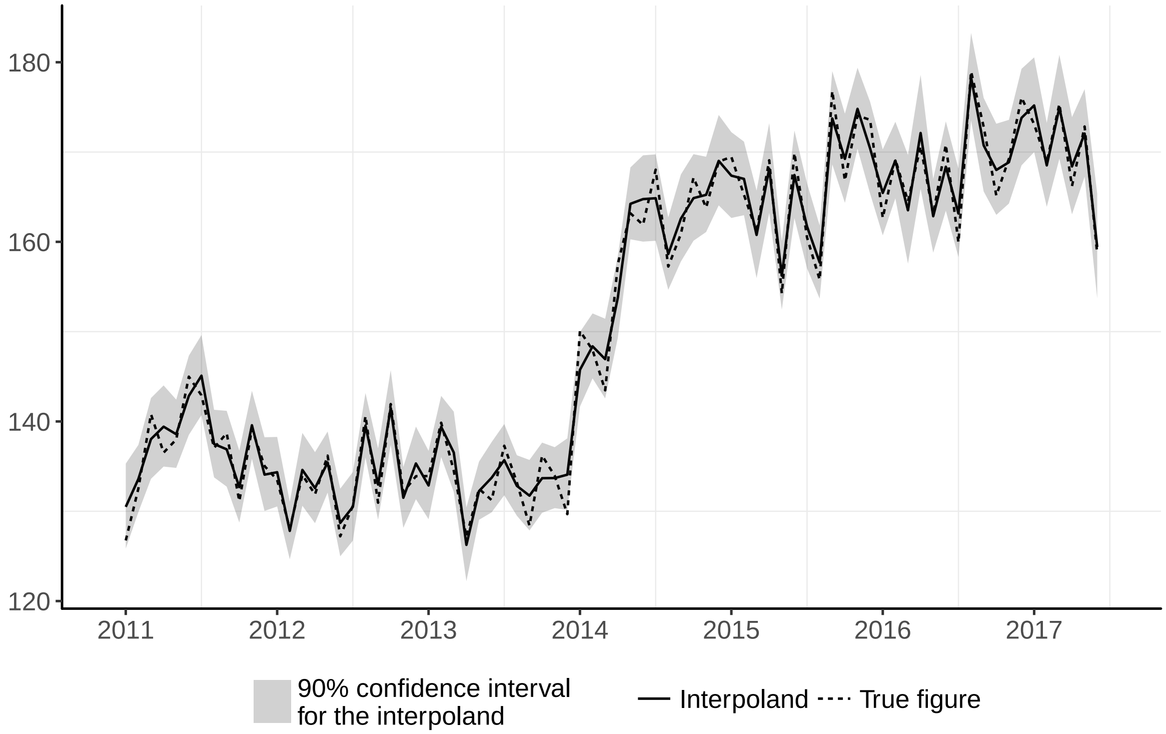


Figure 3.3: Comparison of the interpoland and true figure. Seasonally adjusted monthly estimates. Data from industry 493T495.

increase in the variance of the interpoland's irregular component. Some of the volatility embedded in the monthly VAT figures can thus be retrieved. Both models yield insignificant root mean square approximation errors; it is $3.5e^{-5}$ for the 'univariate' estimation and $4.8e^{-5}$ when estimating correlated components.

There are three important differences compared to the estimation results of the synthetic series. First, while using the covariate for estimation increases the absolute size of the irregular component, it remains smaller than with the synthetic series. Secondly, the estimated observation errors are much larger. It also appears from the variance estimate of the observation errors in table 3.1 that the second stagger is significantly more noisy than the other two. In fact there is a significant improvement in the log likelihood by allowing stagger-specific observation error variances. Thirdly, the estimated stagger biases are large, with a level bias of approximately ten percent for the third stagger at the beginning of the estimation period. For both staggers the bias decreases

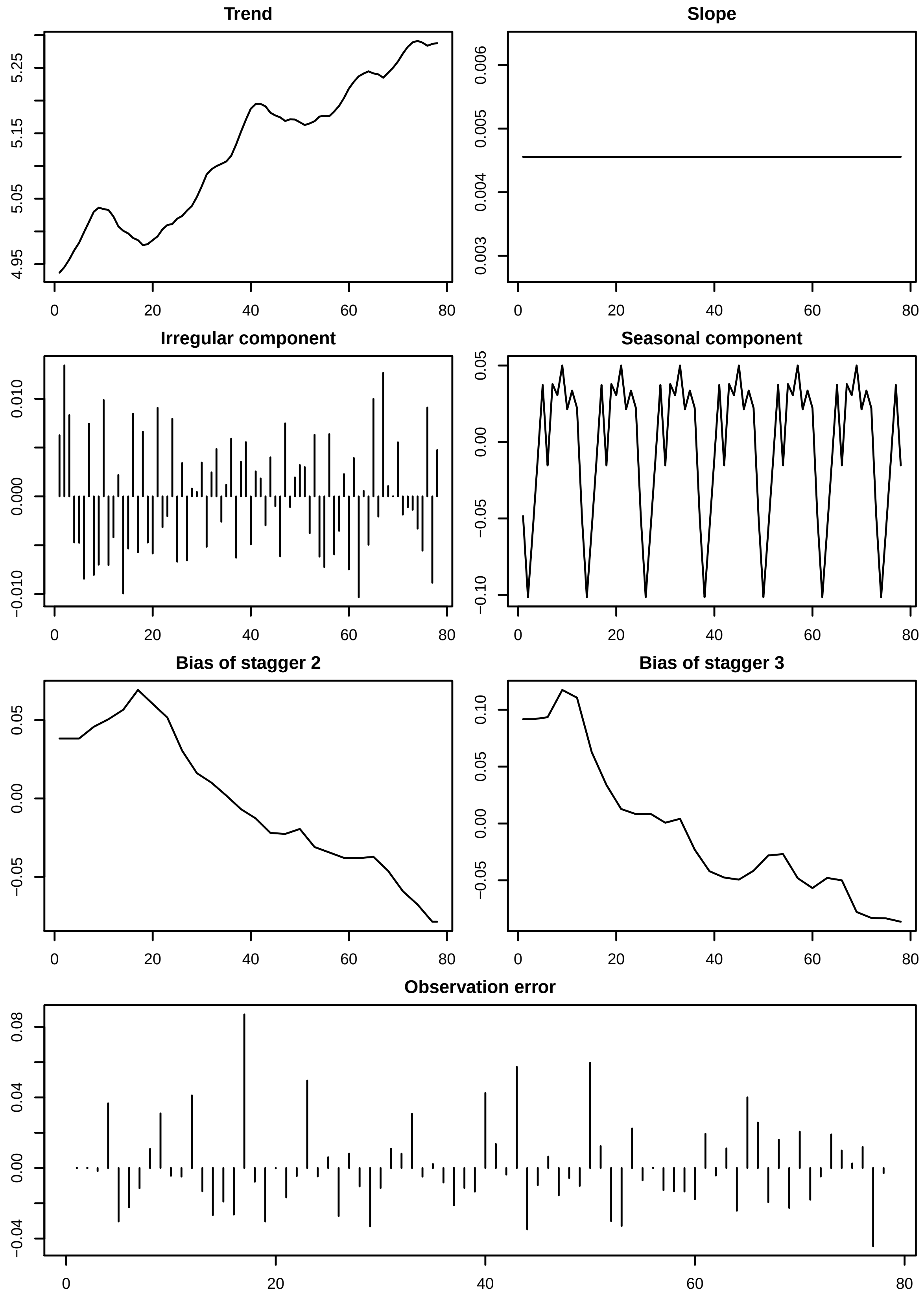


Figure 3.4: Smoothed unobserved components from estimation of the VAT data from industry 493T495. Log units.

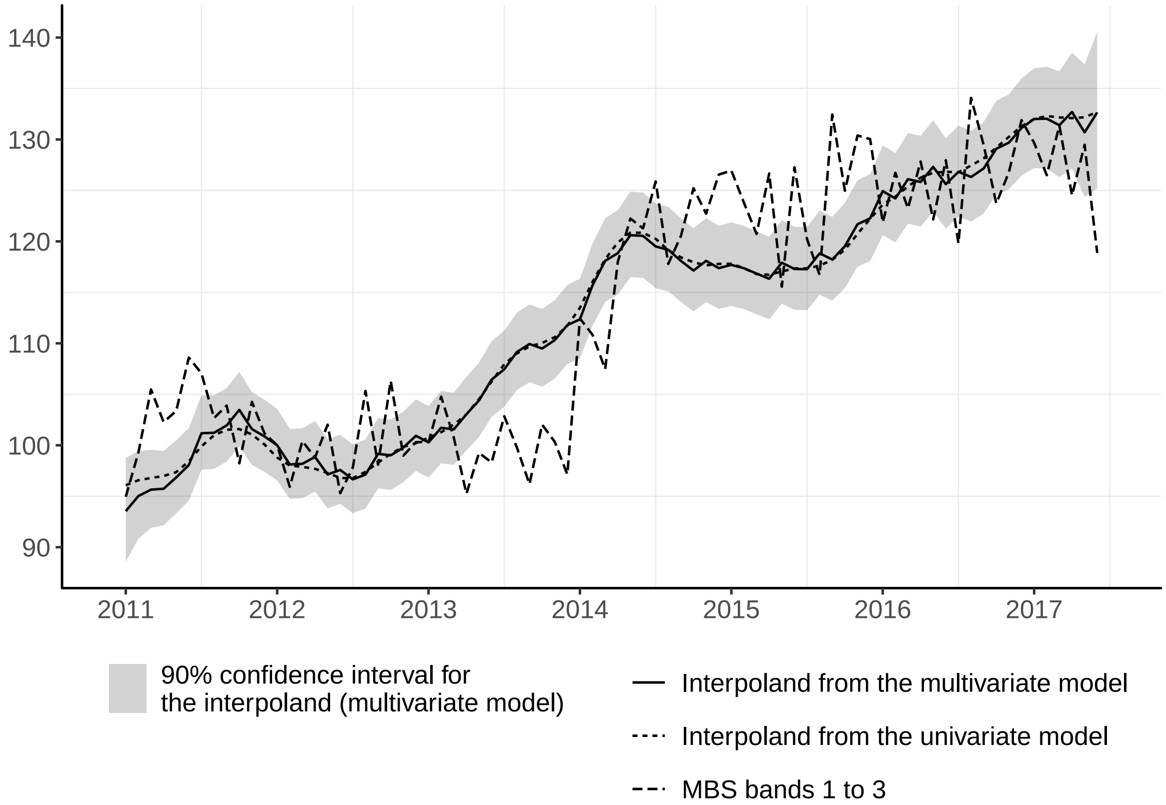


Figure 3.5: Comparison between the seasonally adjusted interpoland and MBS 1-3 figure, index January 2012 = 100. Data from industry 493T495.

over time and becomes negative in the second half of the estimation period.

3.3.3 Comparison with the MBS figures

It is important to see how the VAT seasonally adjusted monthly estimates compare with their MBS counterparts. Figure 3.5 shows the VAT interpoland with its 90% confidence interval alongside the MBS figure. For comparison purpose the MBS series are indexed to the interpolated figure in January 2012.

The trends in the VAT and MBS figures diverge significantly from 2013, although both series represent the same types of businesses. The MBS series is notably subject to a steep level shift at the beginning of 2014 which does not appear in the interpolated data. The upward trend in the VAT data arises sooner than in the MBS figures, and the change is smoother. This steep change in the MBS's trend could be due to the rotating nature of the MBS survey. Interestingly, the growth rates of the two series are inversely related with a correlation coefficient at a monthly frequency of -0.34.

The interpolated series is smoother than the MBS series, which can be explained in two ways. On the one hand, state space estimation typically generates estimates with a lower root mean square log deviation than the underlying 'true' monthly series, as the analysis of synthetic data has illustrated, and measurement errors could exacerbate this feature. On the other hand, survey-based estimates are subject to sampling errors which are likely to create some degree of short-term volatility in the MBS estimates.

3.3.4 Application to all industries

Having illustrated the method with the 'Land transport and transport services via pipelines' industry, aggregate estimates for the whole set of seventy-five industries are now produced. For this the unconstrained multivariate model is used because the results from the case study show that the MBS covariate can be useful for estimating the VAT interpoland. Seasonally adjusted monthly figures in levels are derived for each industry and indexed to January 2011 = 100. They are then aggregated using normalised industry weights calculated as

$$W_{id} = \frac{V_{id}}{\sum_{id=1}^{75} V_{id}}, \tag{3.15}$$

where $id = 1, ..., 75$, indicates the industry and

$$\begin{aligned} V_{id} = &\text{industry } id \text{ gross value added weight } \times \\ &\text{industry } id \text{ turnover share of firms in size bands one to three.} \end{aligned} \tag{3.16}$$

This weighting is necessary because, while the output measure of GDP is produced from gross output indicators, movements in these are combined using net output weights. In particular, a simple aggregation would overstate the importance of the distribution sector whose gross output is measured by margin and not turnover.

The models presented here are logarithmic. For each industry separately this is no obstacle to the calculation of confidence limits, as figures 3.2 and 3.5 illustrates using (3.14). To produce an economy-wide total, however, it is necessary to add the weighted levels, not their logarithms; variances then cannot be similarly aggregated. Therefore indicative confidence limits are produced in the following way. For each industry, upper and lower limits are derived using (3.14) and subsequently indexed using the weighting factor resulting from the indexation of the interpolated series. The approximate deviations of the upper and lower limits from the interpolated figures are referred to as $\tilde{D}_{id,t}^{\text{low}}$ and $\tilde{D}_{id,t}^{\text{high}}$ and used to construct 90% confidence limits for the aggregated interpoland $\hat{x}_t^a$ as

$$\tilde{L}_t^{\text{low}} = \hat{x}_t^a - \sqrt{\sum_{id=1}^{75} (\tilde{D}_{id,t}^{\text{low}})^2 \times W_{id}^2}, \quad \tilde{L}_t^{\text{high}} = \hat{x}_t^a + \sqrt{\sum_{id=1}^{75} (\tilde{D}_{id,t}^{\text{high}})^2 \times W_{id}^2}, \tag{3.17}$$

This approach assumes that the errors in the interpolands are independent across industries. Figure 3.6 shows the aggregated interpolated figures with their 90% indicative confidence interval. The seasonally adjusted aggregate figures from the MBS covering small and medium size businesses, which have been indexed to be equal to the interpolated figure in January 2012, are also plotted.

The VAT-based aggregate turnover figures are clearly less volatile than the aggregated MBS estimates and follow a slightly different trend. Notably, the VAT data point to a slower recovery after the euro area sovereign debt crisis. Fewer than a third of the MBS estimates fall within the 90% confidence bands of the interpolands. The largest discrepancy between the two series arises in January 2014, with a VAT estimate 6.85 percentage points lower than the MBS figure. This implies a discrepancy of about 1.6 percentage points for the output estimates of GDP since the interpolated series cover

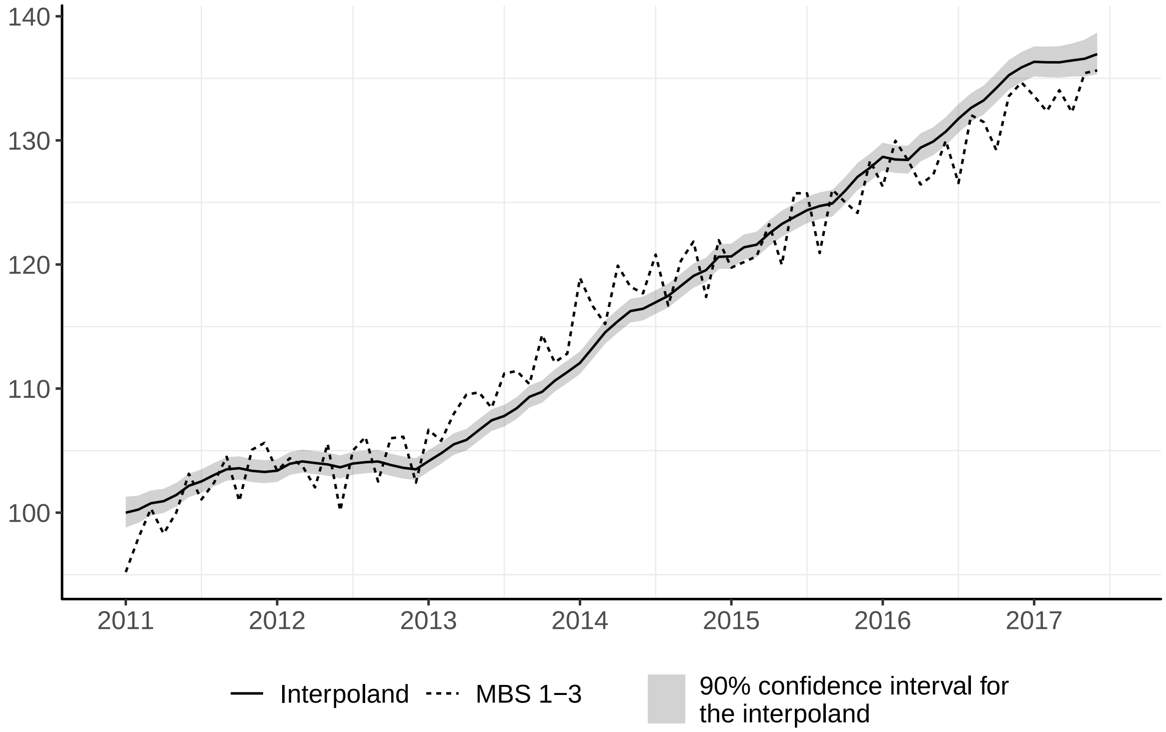


Figure 3.6: Aggregate monthly turnover estimates, seasonally adjusted figures, index January 2012 = 100.

almost a quarter of gross value added in the economy.

The discrepancy between the MBS and VAT estimates could come from the distribution of industries' weights. If industries with high weights exhibit relatively high persistent discrepancies, these could appear in the aggregated figures. However, this is not the case because the log differences between the MBS and VAT-based estimates, for each month across all industries, do not average to zero. Separately, for some industries the MBS and VAT data do not cover exactly the same population of firms. But excluding these industries does not alleviate the divergence between the two series.

3.4 Conclusion

This chapter shows how business turnover figures provided by VAT returns can be used to replace the MBS covering small and medium size businesses. This requires estimating monthly estimates from rolling quarterly figures which is achieved using an unobserved components approach adapted to the particular features of the VAT data. A logarithmic specification is set out to improve estimation, and the resulting nonlinear temporal aggregation constraint is handled with an approximation which proves to be very precise. Avoiding nonlinear techniques facilitates estimation and improves the tractability of the model.

The temporal disaggregation method has been illustrated with the ‘Land transport and transport service via pipelines’ industry as a case study. Empirical results using a synthetic series suggest that the approach is efficient at disaggregating temporally overlapping seasonal data, but the measurement errors in the VAT series complicate the identification of the monthly movements in the interpolands. This issue is alleviated partially by using the MBS for the largest businesses as an indicator series in a bivariate framework.

Finally, VAT-based aggregate figures of turnover covering approximately a quarter of gross value added in the economy are derived. While the VAT data can be used to replace part of the MBS there is, nevertheless, a difference in their profile. Replacement of survey data by administrative data may, thus, lead to some rewriting of economic history.

Chapter 4

Nowcasting in the presence of large measurement errors and revisions

4.1 Introduction

Chapter 3 presents a flexible unobserved component model in a state space framework for deriving monthly output figures from rolling quarterly VAT data. In this chapter this approach is extended to tackle another feature of these data: the delay and highly noisy nature of the early figures. The main contribution of this chapter lies in the presentation and illustration of a cleaning method suited to capture non-Gaussian features in the distribution of the measurement errors such as asymmetric tails and extreme observations.

ONS publish monthly GDP figures about two months after the end of the reference period, but at this time only a relatively small proportion of the VAT data are available. VAT returns begin accruing shortly after the end of the reference period but take several months to be complete. Therefore, an important question is whether one can derive precise estimates of monthly output from the data available two months after the end of the reference period.

By nature the VAT-based turnover data become more precise as more respondents fill their VAT returns, and early releases can be subject to biases. It is possible to capture the revisions across releases as well as their differing levels of measurement errors by modelling them together. The use of an unobserved component model for temporal disaggregation lends itself naturally to this kind of extension. The model presented in the previous chapter is a bivariate unobserved components model where a covariate is used to improve VAT-based monthly output, precisely the MBS for large businesses. This chapter shows how this model can be extended with successive releases, or snapshots, of the VAT data, in order to capture the revision pattern and thus forecast monthly output using early data.

But this nowcasting exercise is complicated by the extremely noisy nature of the early figures. Applying the Kalman filter and smoother to the original VAT releases, even when accounting for the heterogeneity across releases, produces very erratic results. It is necessary to clean the data from very large measurement errors before estimating the nowcasting model.

Two approaches are explored for cleaning the VAT data prior to nowcasting. The first approach, which is the common strategy when using state space models, consists of carrying t-tests on standardised disturbances to detect outlying observations. These observations are discarded and replaced with missing values and the model is re-estimated until no new outliers are detected. The Kalman filter - the recursive algorithm used for estimating state space models - handles missing values straightforwardly.

But as is illustrated in this chapter, estimation with original figures for detecting outliers typically yields residuals which fail to be normally distributed. And the state space techniques presented in chapter 3 relies on the data being conditionally normally distributed. To model openly for non-Gaussian features in the data a second cleaning strategy relying on score driven methods introduced by Creal et al. (2013)

and Harvey (2013) is exploited. This method deals with large measurement errors without systematically discarding them; instead it uses a complex weighting scheme for downweighting large prediction errors which is linked to their conditional distribution.

The uncertainty surrounding the nature of the noise affecting the VAT figures implies that combining both cleaning approaches could be beneficial. Therefore, both approaches are compared with a third strategy consisting of nowcasting separately with both cleaning approaches and averaging the resulting nowcasts.

The VAT series begin in March 2011 for the seventy-five industries for which they are available, but vintages are observable only from January 2012. The last month in which all releases are observable is September 2019; after that the data are gradually missing. The missing observations at the beginning and end of the sample are not problematic because the filtering techniques used for estimation below handle them straightforwardly by producing recursively optimal forecasts for missing values.

The next section investigates the extend of the revisions across VAT releases and show that they are biased significantly. Using a simple trimming rule it also illustrates the magnitude of the noise affecting the data and its asymmetric nature. Accordingly, section 4.3 shows how the bivariate model presented in chapter 3 can be extended with additional components for capturing these revisions, and section 4.4 sets out and illustrates the two approaches considered for cleaning the data of extreme measurement errors before nowcasting. Using a set Monte Carlo experiments it is shown that the score driven cleaning method outperforms the standard approach of discarding outliers through t-tests when measurement errors follow non-Gaussian processes.

The nowcasting exercise is carried out in pseudo real-time to mimic a publication schedule of a National Statistical Institute, and section 4.5 discusses the results. These show that the score driven cleaning approach yields smaller revisions to the monthly VAT-based output nowcasts than the t-test method, and that averaging both approaches does not help. Secondly, despite producing early estimates which tends to be revised downward, the VAT-based nowcasts can indicate timely an economic recession, as is illustrated with the Covid-19 pandemic period. Finally, although the VAT returns and MBS provide similar picture of monthly output changes overtime, there can be some persistent discrepancy in their levels, confirming the finding of chapter 3.

4.2 Descriptive analysis of the revisions and noise

Each month a new series is released which shows a VAT-based quarterly turnover observation for a new month as well as revised values for past months. The most recent observation of this vector is stored in a vector showing the first releases over time; the second observation is stored in a vector showing the second releases over time, and so on. These successive snapshots of a same variable are usually represented with a 'revision triangle' like Figure 4.1. The log quarterly turnover for a given industry in month t, and observable in release i, is defined as $y_{i,t}$. It is assumed that the eleventh release is the final release such that $i = 1, ..., 11$. Observations are still subject to some revisions after the eleventh release but only on a very small scale. The VAT figures are observable consistently starting from two months after the end of the reference period; hence the first release ($i = 1$) has a two-month lag.

The VAT releases are weighted by ONS to account for the proportion of missing respondents, but they remain biased and differ in their level of measurement errors. To illustrate this fact a preliminary analysis is carried out on the revisions between the early releases and the last release, twelve months after the reference period at $i = 11$. These ten revisions are given by

$$rev^j_{i,t} = y^j_{i,t} - y^j_{11,t}, \quad i = 1, ..., 10, \tag{4.1}$$

$$
\begin{bmatrix}
y_{11,1} & y_{10,1} & y_{9,1} & \cdots & y_{2,1} & y_{1,1} \\
y_{11,2} & y_{10,2} & y_{9,2} & \cdots & y_{2,2} & y^{1}_{i,2} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
y_{11,t} & y_{10,t} & y_{9,t} & \cdots & y_{2,t} & y_{1,t} \\
 & y_{10,t+1} & y_{9,t+1} & \cdots & y_{2,t+1} & y_{1,t+1} \\
 & & \ddots & \vdots & \vdots & \vdots \\
 & & & y_{3,t+8} & y_{2,t+8} & y_{1,t+8} \\
 & & & & y_{2,t+9} & y_{1,t+9} \\
 & & & & & y_{1,t+10}
\end{bmatrix}
$$

Figure 4.1: Illustration of the revision triangle of the VAT data for a given industry. Columns and first subscripts indicate the releases while rows and second subscripts indicate the month the figures relate to. The maturity of the data increases from left to right and their timeliness increases from top to bottom. This matrix show the data observable at $t + 10$.

where here $j = 1, ..., 75$ indicates the industry index.

The first two rows of table 4.1 show the aggregate mean revision and its standard error for each early release. Revisions are computed at industry level using (4.1) and weighted using the industries' shares in total gross value added and shares of turnover covered by firms in size bands one to three. Hence the weighted mean is

$$\overline{rev}_i = \frac{\sum_{j=1}^{N} \sum_{t=1}^{T} w_{j,t} rev^{j}_{i,t}}{\sum_{i=1}^{N} \sum_{t=1}^{T} w_{j,t}}, \tag{4.2}$$

where

$w_{j,t} =$ Share of turnover covered by firms in size bands 1 to 3 in industry i in month $t\times$

Contribution of industry i in total gross value added in month t.

Gatz and Smith (1995) discuss several methods to compute the standard error of the mean when the data are weighted. Here the simplest approach is adopted and the standard error are computed as

$$se_i = \sqrt{\frac{1}{NT} \frac{\sum_{j}^{N} \sum_{t}^{T} w_{j,t} (rev^{j}_{i,t} - \overline{rev}^{j}_{i})^2}{\sum_{i}^{N} \sum_{t}^{T} w_{j,t}}}. \tag{4.3}$$

Table 4.1: Weighted mean and standard error of revisions at each maturity level with the implied z-scores.

Original data										
j	1	2	3	4	5	6	7	8	9	10
Mean	0.19	-1.27	-1.02	-0.85	-0.69	-0.58	-0.37	-0.17	-0.13	-0.08
S.E.	0.31	0.19	0.13	0.11	0.1	0.1	0.07	0.04	0.04	0.03
Z-score	0.59	-6.63	-7.78	-7.49	-6.97	-5.72	-5.27	-4.15	-2.84	-2.69

Trimmed data										
j	1	2	3	4	5	6	7	8	9	10
Mean	2.34	-0.5	-0.44	-0.39	-0.32	-0.22	-0.17	-0.11	-0.06	-0.05
S.E.	0.09	0.04	0.04	0.04	0.03	0.03	0.03	0.02	0.03	0.03
Z-score	25.11	-11.61	-11.05	-10.78	-10.26	-6.47	-5.36	-4.39	-2.11	-1.7

Note: Weighted mean of the revisions $rev_{i,t}^{j} = y_{i,t}^{j} - y_{11,t}^{j}, \quad i = 1, ..., 10$, using equation (4.2) with GVA weights. Standard error of the mean using equation (4.3)

To get an indication of the potential biases present in the early releases the third row of table 4.1 shows z-scores for weighted means. In aggregate, revisions are biased at all maturities and their standard errors typically decrease with maturity. As more data become available the releases become more precise and less volatile with a decreasing bias.

The presence of these biases in the VAT data requires a flexible nowcasting approach capable of capturing large and potentially dynamic biases. The next section shows how the bivariate model presented in chapter 3 may be extended with an additional set of unobserved components aimed at capturing these specific features of the data.

4.3 Nowcasting approach

The early VAT data are incomplete because most firms submit their VAT returns with a delay. As more data accrue over time, the VAT figures are revised and their precision improves. To produce a timely monthly estimate of turnover from the VAT data it is necessary to use the earliest release, available with a two-month lag, and since the data are revised over time, the model should also take into account the accrual of new information and update past estimates. Both tasks can be achieved by modelling the eleven VAT releases together in a multivariate framework.

Building on the bivariate model presented in section 3.2, each VAT release is modelled as

$$\begin{aligned} y_{i,t} &= \tilde{y}_{1,t} + r_{i,t}, \\ &= \ln 3 + \frac{1}{3}x_{1,t} + \frac{1}{3}x_{1,t-1} + \frac{1}{3}x_{1,t-2} + \gamma_{1,t} + \beta_1 h^a_{1,t} + b^{(j)}_{1,t} + r_{i,t}, \end{aligned} \tag{4.4}$$

where $\tilde{y}_{1,t}$ is the VAT-based quarterly turnover *signal* representing the economically relevant information common to all release, while $r_{i,t}$ is the observation error affecting release i. This measurement error is modelled as the sum of a bias specific to each early release $c_{i,t}$, $i = 1, ..., 10$ and white noise $\epsilon_{i,t}$:

$$r_{i,t} = c_{i,t} + \epsilon_{i,t}, \tag{4.5}$$

where the biases take the form of a random walk:

$$c_{i,t} = c_{i,t-1} + \tau_{i,t}, \quad \tau_{i,t} \sim N(0, \sigma_{\tau,i}), \tag{4.6}$$

and $\epsilon_{i,t}$ is white noise independent across releases:

$$\epsilon_{i,t} \sim N(0, \sigma_{\epsilon,i}). \tag{4.7}$$

By construction the last release in assumed to be unbiased. Importantly, the presence of a measurement error in the final release means that the ‘true’ quarterly turnover might never be observed; this is important because a substantial amount of noise remains in the final release.

For clarity the covariate is denoted as z_t and is modelled as

$$z_t = x_{2,t} + \gamma_{2,t} + \beta_2 h^a_{2,t}. \tag{4.8}$$

The covariate is as timely as the first release and is used to improve the measure of the seasonally adjusted VAT-based monthly turnover $x_{1,t}$. This is achieved by modelling seasonally adjusted figures, $x_{1,t}$ and $x_{2,t}$ together in a bivariate local linear trend model presented in 3.2.4.

State space representation and estimation

The observation and state equations can be cast together in state space form as

$$\begin{aligned} y_t &= Z_t \alpha_t, \\ \alpha_{t+1} &= T\alpha_t + R\eta_t, \qquad \eta_t \sim \mathrm{N}(0, Q), \\ \alpha_1 &\sim \mathrm{N}(a_1, P_1), \end{aligned} \tag{4.9}$$

where $y_t = (y_{1,t} - \log 3, ..., y_{11,t} - \log 3, z_t)'$ and $\alpha_t = (\alpha'_{1,t}, \alpha'_{2,t})'$; $\alpha_{1,t} = (\mu_{1,t}, \mu_{1,t-1}, \mu_{1,t-2}, \nu_{1,t}, \gamma_{1,t}, ..., \gamma_{1,t-8}, e_{1,t}, e_{1,t-1}, e_{1,t-2}, \beta_1, b^1_{1,t}, b^2_{1,t}, b^3_{1,t}, c_{2,t}, ..., c_{10,t}, \epsilon_{1,t}, ..., \epsilon_{11,t})'$, $\alpha_{2,t} = (\mu_{2,t}, \nu_{2,t}, \gamma_{2,t}, ..., \gamma_{2,t-10}, e_{2,t}, \beta_2)'$. Appendix D.2 shows the full matrix representation of the model.

The unknown variance parameters are estimated via maximum likelihood, where the log likelihood function is evaluated with the prediction error decomposition using the Kalman filter's output. The Kalman filter is a recursion initialised with the mean vector a_1 and covariance matrix P_1 of the initial state vector which are unknown. The trends, slopes, seasonal effects, Easter effects, stagger biases and release biases are initialised with an exact diffuse initialisation, that is to say with arbitrary means and infinite variances, while the stationary states (the irregular components and noise) are initialised with zero means and their unconditional variances.

While the Kalman filter yields optimal predictions of the state vector, which is suited to estimation and forecasting, the estimate of the state vector and its error variance given the entire sample is given by the Kalman smoother. The Kalman smoother is a backward recursion which makes use of the Kalman filter output when the parameters are set to their maximum likelihood estimates. The interpolands in levels are constructed from the smoothed estimates of the states as

$$\exp(\hat{x}_{1,t}) = \exp(\hat{\mu}_{1,t} + \hat{e}_{1,t}), \qquad t = 1, ..., T. \tag{4.10}$$

The Kalman filter and smoother algorithms with exact diffuse initialisation of Durbin and Koopman (2012) and the associated log likelihood function is used.

When revisions are gradually missing towards the end of the series, as is illustrated with the revision triangle in figure 4.1, the dimension of the observation vector y_t is adapted accordingly. Hence, the information from early revisions can be used although not all revisions are available. The Kalman filter produces optimal forecasts for these missing values.

4.4 Cleaning strategies

Estimating the nowcasting model presented in section (4.3) using the original VAT data produces poor results. This is because the VAT releases are subject to extremely large measurement errors; these are illustrated for the aggregated data in the first panel of figure 4.8. Some of these errors are likely to come from firms reporting turnover figures with decimal errors. Indeed, since only small and medium size businesses are analysed, a single business cannot have an important impact on the industry aggregate unless it is a reporting error. This also means that outlying observations are most often positive because abnormally low figures do not affect the aggregate very much.

To improve the VAT-based monthly output nowcasts it is necessary to clean the data from these very large measurement errors beforehand. For this two separate approaches are explored because there is a limited understanding regarding the behaviour of the measurement affecting the VAT data.

4.4.1 Method 1: Sequentially discarding outlying observations

The first cleaning method consists of carrying out t-tests sequentially on the standardised observation errors. These are given by

$$\hat{\epsilon}^s_{i,t} = \hat{\epsilon}_{i,t}/\sqrt{H_{i,t}}, \quad i = 1, ..., 11, \tag{4.11}$$

where $H_{i,t}$ is the estimated variance. To derive the standardised observation errors the bivariate model (3.11) is estimated at industry level for each release, where the VAT data is modelled together with the covariate. After estimation the Kalman smoother is used to retrieve the smoothed observation errors and their estimated variance.

Each standardised observation error can be tested for significance using a two-tailed t-test with the null hypothesis that the corresponding observation is not an outlier. Outlying observations thus detected are discarded and replaced with missing values. Next the bivariate model is estimated again and a new set of t-tests are carried out on the standardised disturbances. This process is iterated until no standardised disturbances reach the threshold indicating outliers.

It is common to choose a confidence interval of 95% which is associated with critical values of ±1.96. However, this is a multiple testing problem and too many outliers would be detected by chance. Because of the large number of industries studied, analysing each outlying observation individually to decide whether or not it should be excluded is not feasible. Therefore the confidence interval is set to a very conservative range of 99.9% which is associated with critical values of ±3.3.

Outliers in these errors indicate observations which cannot be explained by the model. These can be due to extremely large measurement errors or misspecification in the model. Setting a high critical value increases the confidence that the outlying observation detected is a measurement error and not a structural break or one-off economically relevant event. In addition, since the data are rolling quarterly aggregates, an outlier carrying economic meaning should appear in three consecutive observation errors.

An important drawback of the method relying on t-tests is that it requires choosing critical values indicating outlying observations, and this inevitably generates a trade-off between discarding too many observations and taking the risk of not detecting large measurement errors. It also assumes that measurement errors are normally distributed. However, as discussed above, measurement errors in the VAT data are most often positive; this could be one source of non-normality. The second cleaning method proposed below deals with outlying observations without discarding them completely and can handle non-Gaussian features.

4.4.2 Method 2: Downweighting observations with a score driven model

Creal et al. (2013) and Harvey (2013) derive predictive filters based on the score of the conditional (or predictive) likelihood, which can arise from a wide range of distributions. These score driven models provide an alternative approach for cleaning where outlying observations are downweighted instead of being discarded completely. The downweighting scheme depends on the magnitude of the prediction errors and their estimated distribution. Using this approach it is possible to model openly the non-Gaussian features in the distribution of the measurement errors. In a cleaning context this method is used to derive pseudo observations where measurement errors are downweighted depending on their magnitude.

A general location model

The model is based on the general location Dynamic Conditional Score (DCS) model of Harvey (2013) and Harvey and Luati (2014). The model is univariate; each release is modelled separately. The general location model is

$$\begin{aligned} y_{i,t} &= Z_t a_t + v_t, \\ a_{t+1} &= T a_t + \kappa u_t, \end{aligned} \tag{4.12}$$

for releases $i = 1, ..., 11$, where κ is a vector of unknown parameters and u_t is the scaled score vector defined as

$$u_t = \frac{\partial \ell_t}{\partial a_t} . s_t^{-1}, \tag{4.13}$$

where $\ell_t = \ln p(y_{i,t}|Z_t a_t; \Theta)$ is the predictive log likelihood with Θ a vector of fixed parameters. The scaling factor s_t is typically related to the information matrix. Conditional on past observations $y_{i,t}$ follows a generalised asymmetric student-t (AST) distribution of Zhu and Galbraith (2010). Hence

$$\begin{aligned} \ell_t =& - \ln\sigma - \frac{\nu_1 + 1}{2} \ln\Big[1 + \frac{1}{\nu_1}\Big(\frac{y_{i,t} - Z a_t}{2\alpha\sigma K(\nu_1)}\Big)^2\Big] \; 1(y_{i,t} \leq Z_t a_t) \\ & - \frac{\nu_2 + 1}{2} \ln\Big[1 + \frac{1}{\nu_2}\Big(\frac{y_{i,t} - Z_t a_t}{2(1-\alpha)\sigma K(\nu_2)}\Big)^2\Big] \; 1(y_{i,t} > Z_t a_t) \end{aligned} \tag{4.14}$$

where σ is the scale parameter, α is the skewness parameter which can take values in $[0, 1]$, ν_1 and ν_2 are respectively the left and right tail parameters, $K(\nu) = \Gamma((\nu + 1)/2)/(\sqrt{\nu\pi}\Gamma(\nu/2))$ ($\Gamma(.)$ being the Gamma function) and $1(x)$ is an indicator variable equal to one if statement x is true and zero otherwise. The distribution is skewed to the right if $\alpha < 1/2$ and to the left if $\alpha > 1/2$.

It is possible to recover well-known Student's t-distributions by restricting and redefining the parameters of the AST. With $\nu_1 = \nu_2$ (symmetric tails), $\alpha = 0.5$ (no skewness) and redefining the scale as $\sigma = \sigma K(\nu)$, the AST reduces to the Student's t-distribution. With $\nu_1 = \nu_2$ and redefining the shape and scale parameters as $\alpha = 1/(1+\gamma^2)$ and $\sigma = (\gamma + 1/\gamma)\sigma K(\nu)/2$ the AST is equivalent to the skewed t-distribution of Fernández and Steel (1996). The skewed Student's t-distributions of Azzalini and Capitanio (2003) and Gomez et al. (2007) cannot be recovered by simple reparametrisation.

An attractive feature of (4.14) is that each tail can have distinct rates of decay captured by the tail parameters. This is particularly useful for the VAT data because extreme measurement errors are most often positive. Hence the right side of the distribution is likely to exhibit a fat tail but not necessarily the left side.

The score vector is

$$\begin{aligned}\frac{\partial \ell_t}{\partial a_t} = & \frac{\nu_1 + 1}{1 + \frac{1}{\nu_1}\left(\frac{v_t}{2\alpha\sigma K(\nu_1)}\right)^2} \cdot \frac{v_t}{\nu_1(2\alpha\sigma K(\nu_1))^2} \; 1(y_{i,t} \leq Z_t a_t) \\ & + \frac{\nu_2 + 1}{1 + \frac{1}{\nu_2}\left(\frac{v_t}{2(1-\alpha)\sigma K(\nu_2)}\right)^2} \cdot \frac{v_t}{\nu_2(2(1-\alpha)\sigma K(\nu_2))^2} \; 1(y_{i,t} > Z_t a_t).\end{aligned} \tag{4.15}$$

and the scaling factor is

$$\begin{aligned}s_t = & \frac{\nu_1 + 1}{\nu_1(2\alpha\sigma K(\nu_1))^2} \; 1(y_{i,t} \leq Z_t a_t) \\ & + \frac{\nu_2 + 1}{\nu_2(2(1-\alpha)\sigma K(\nu_2))^2} \; 1(y_{i,t} > Z_t a_t).\end{aligned} \tag{4.16}$$

The scaled score (4.13) thus takes the form

$$\begin{aligned}u_t = & \frac{v_t}{1 + \frac{v_t^2}{\nu_1(2\alpha\sigma K(\nu_1))^2}} \; 1(y_{i,t} \leq Z_t a_t) \\ & + \frac{v_t}{1 + \frac{v_t^2}{\nu_2(2(1-\alpha)\sigma K(\nu_2))^2}} \; 1(y_{i,t} \leq Z_t a_t).\end{aligned} \tag{4.17}$$

The form of (4.17) is close to the scaled score of Harvey (2013) (page 96), where the scaling factor diverges slightly from the information quantity (therefore it diverges from the location scaling factor of the next two chapters). Notably $u_t = v_t$ in the Gaussian case when $\nu_1, \nu_2 \to \infty$ and $\alpha = 1/2$. The distance between the prediction error and the scaled score thus indicates the degree of nonlinear weighting and divergence from the Gaussian model.

Figure 4.2 shows the scaled score as a function of the prediction error. Low tail parameters downweight the effect of large prediction errors, while the scaled score reduces to the prediction error when tail parameters tend to infinity, which is the behaviour intended with this specific choice of scaling factor. Note that low tail parameters do not induce an overweighting of small prediction errors, as is typically the case when downweighting the score with its information quantity (see notably chapter 4 and Delle-Monache and Petrella (2017)). The absence of overweighting is consistent with a cleaning application where observation errors are downweighted whereas common observations remains unaffected.

To compare the score driven with the typical approach of discarding outliers, 4.2 also plots the response function driving the latent states when the distribution of the prediction error is constrained to be Gaussian (in this case, score driven models for location parameters can be equivalent to the Kalman filter at steady state, see Buccheri et al. (2021)) and outliers are discarded with t-tests. This response function is an approximation of the response function implied by Gaussian state space models; it is only an approximation because observation errors in state space models are derived using the Kalman smoother which, unlike the Kalman filter, makes use of the entire sample. While the use of t-tests generates a discontinuity in the response function, the scaled score downweights prediction errors gradually depending on their magnitude.

The benefit of the asymmetric Student's t-distribution is illustrated with a low tail

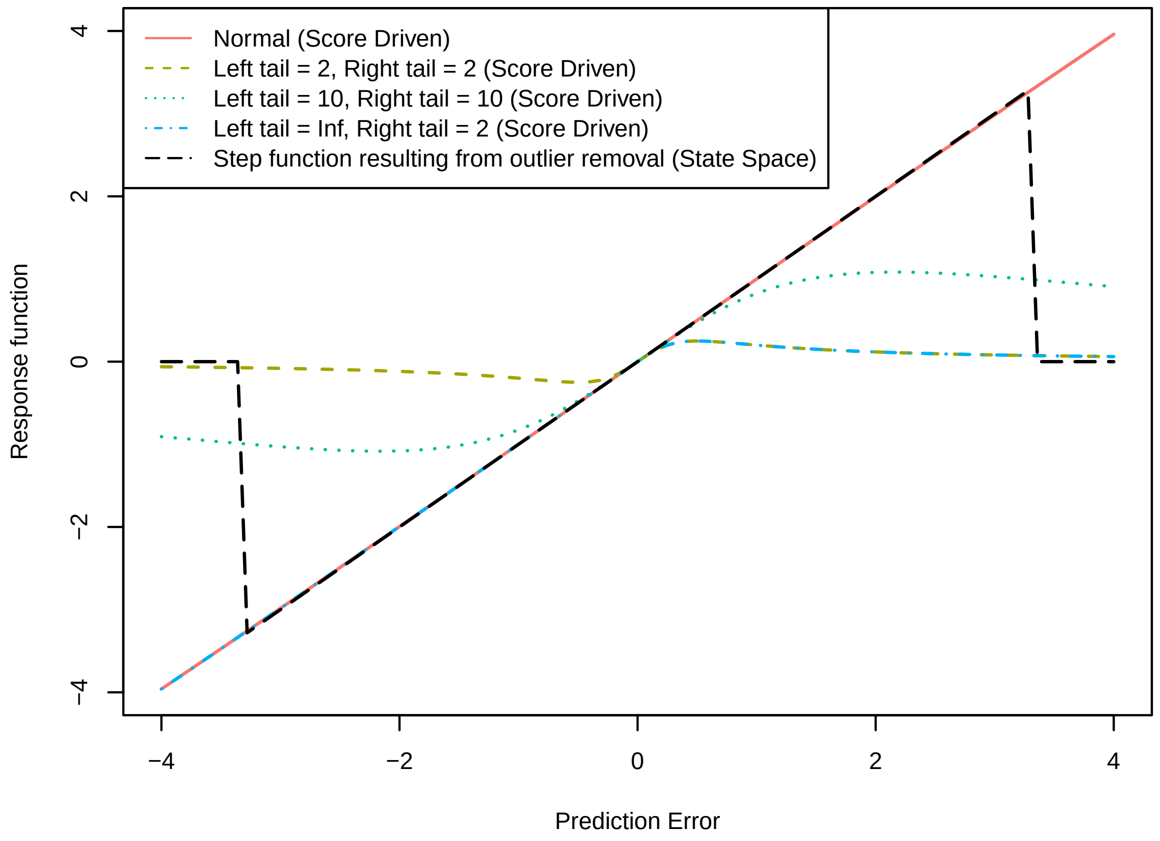


Figure 4.2: Illustration of the response function driving the unobserved components in score driven and state space models. In a score driven model the response function is the scaled score u_t; In a state space model discarding outliers yields a linear response function for prediction errors not detected as outliers and zero otherwise (critical value of +-3.3). The scale parameter is set to one ($\sigma = 1$).

parameters for negative prediction errors (dotted blue line). This set of parameters yield a larger downweighting of positive errors but a linear response for negative prediction errors. This behaviour of the scaled score with asymmetric tails is especially fitted for the VAT data because observation errors tend to be positive only. Finally, although low tail parameters never yield a response function larger than the prediction error, this is not a limitation when the model is estimated; indeed low tail parameters typically yield a smaller scale parameter which in turn gives a greater scaled score (see equation (4.17)).

The recursion is initialised with the initial state vector a_1 which is typically unknown. While with the standard Kalman filter one can resort to a diffuse initialisation, this is not possible with score driven models. Instead the initial state vector is estimated via maximum likelihood along the other unknown parameters. Specifically, the vector of unknown fixed parameter $\Theta = (\nu_1, \nu_2, \alpha, \sigma, \kappa, a_1')'$ is estimated by numerical maximisation of the log likelihood as

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{t=3}^{N} \ell_t.$$

To improve estimation an analogous Gaussian unobserved components model is estimated beforehand and the resulting initial smoothed state vector is used as starting values for a_1.

Smoothing

It is not possible to generate directly smoothed unobserved components with score driven techniques, but pseudo observations where non-Gaussian features have been alleviated can be retrieved. This is what Caivano et al. (2016) propose through an iterative algorithm which makes use of a Gaussian unobserved components model. This method consists of the following iterative scheme

1: Estimate the score driven model (4.12) with the original data $y_{i,t}$;

2: Generate pseudo observations as $\tilde{y}_{i,t} = Z_t a_t + u_t$;

3: Estimate the unobserved components model with the pseudo observations $\tilde{y}_{i,t}$ and retrieve the smoothed estimates $\hat{\alpha}_t$;

4: Set $v_t = y_{i,t} - Z_t \hat{\alpha}_t$ in (4.15) which yields new values of u_t through (4.17);

5: Generate pseudo observations as $\tilde{y}_{i,t} = Z_t \hat{\alpha}_t + u_t$;

6: Iterate steps 3 to 5 until convergence.

If the estimated distribution is Gaussian $u_t = v_t$, which implies that the pseudo observations constructed from the score driven recursion are equal to the original observations. In this case the smoothing algorithm is reduced to the Kalman smoother. Once the algorithm has converged the pseudo observations $\tilde{y}_{i,t}$ are used for estimating model (4.9).

The pseudo observations thus derived are not free of measurement errors, but the remaining noise should be normally distributed and can therefore be captured effectively with standard state space techniques.

4.4.3 Monte Carlo experiments

This section investigates the stability and efficiency of the score driven method presented above when applied to a relatively short sample (approximately 100 observations) while comparing different level of restrictions on the model generating the prediction errors. The most efficient model specification is then used in another experiment designed to analyse the effectiveness of the score driven cleaning approach compared to the more common strategy of discarding outliers through t-tests.

On the stability and efficiency of the score driven method

For studying the stability of the score driven estimation, synthetic data are generated using model (4.12) with different specifications for the distribution of the prediction error v_t. Specifically, in each specification the vector of prediction error v_t is generated using an asymmetric student-t distribution with different levels of restriction on its parameters. This is useful to understand if a greater flexibility in the distribution of the prediction error is worth the increase in the number of parameters to estimate, which can be difficult with a relatively short sample. The efficiency of the score driven estimation is investigated using the root mean square error between the estimated pseudo observation $Z_t\hat{a}_t + \hat{u}_t$ and its true value (stored when generating the data). The results are calculated from 100 simulations for each model specification.

The different model restrictions tested in the Monte Carlo experiment are shown in table 4.2. The most flexible model can be skewed and have different tail parameters. Restricting the skewness to zero but allowing for distinct tail parameters yields to a Student-t model with distinct tail parameters. Conversely, restricting the tail parameters to be identical but allowing the distribution to be skewed yields a Skewed student-t model. Restricting the tail parameters to be identical and constraining the skewness to zero yields the Student-t model, while constraining the tail parameters to very large or infinite values but letting the skewness parameter vary yields a Skew Normal distribution. Finally, restricting the tail parameters to very large or infinite values and fixing the skewness to zero gives the Normal distribution.

Table 4.2: Model specifications for the prediction error in the Monte Carlo experiments. All specifications are derived from the Asymmetric Student-t distribution with different degrees of restriction on the parameters.

Model	Location	Left tail	Right tail	Scale	Shape
FT Skew Asy.	$\mu > 0$	$\nu_1 > 0$	$\nu_1 > 0$	$\sigma > 0$	$1 > \alpha > 0$
FT Asy.	$\mu > 0$	$\nu_1 > 0$	$\nu_1 > 0$	$\sigma > 0$	$\alpha = 0.5$
FT Skew	$\mu > 0$	$\nu > 0$		$\sigma > 0$	$1 > \alpha > 0$
FT	$\mu > 0$	$\nu > 0$		$\sigma > 0$	$\alpha = 0.5$
Skew	$\mu > 0$	$\nu = \infty$		$\sigma > 0$	$1 > \alpha > 0$
Gaussian	$\mu > 0$	$\nu = \infty$		$\sigma > 0$	$\alpha = 0.5$

Table 4.3: Simulation results with true parameters set to $\boldsymbol{\alpha = 0.5 \quad \nu_2 = 1}$; $\sigma = 5 \quad \nu_1 = 250 \quad k_1 = 0.5$

Model	AIC	ν_1	ν_2	k_1	RMSE	σ	α
FT Skew Asy.	-375.87	157.96	1.12	1.13	0.76	4.14	0.42
FT Asy.	-373.91	146.72	0.93	0.98	0.66	4.2	0.5
FT Skew	-357.25	1.28	1.28	2.2	1.26	3.47	0.44
FT	-356.7	1.1	1.1	2.06	1.19	3.51	0.5
Skew	-191.2	250	250	0.12	4.36	171.37	0.03
Gaussian	-116.94	250	250	0.13	4.42	185.91	0.5

Tables 4.3 to 4.11 show the results. Overall having distinct tail parameters is important; however, constraining the skewness to zero yields to better results. Next a

Table 4.4: Simulation results with true parameters set to $\boldsymbol{\alpha = 0.4} \quad \boldsymbol{\nu_2 = 1}$; $\sigma = 5 \quad \nu_1 = 250 \quad k_1 = 0.5$

Model	AIC	ν_1	ν_2	k_1	RMSE	σ	α
FT Skew Asy.	-361.77	167.55	0.96	1.21	0.69	3.71	0.35
FT Asy.	-356.27	174.73	0.76	0.77	0.51	3.94	0.5
FT Skew	-344.93	0.83	0.83	2.2	1.05	2.82	0.48
FT	-346.64	0.81	0.81	2.27	1.06	2.73	0.5
Skew	-147.93	250	250	0.1	4.35	31.58	0.02
Gaussian	-62.89	250	250	0.1	4.47	53.65	0.5

Table 4.5: Simulation results with true parameters set to $\boldsymbol{\alpha = 0.3} \quad \boldsymbol{\nu_2 = 1}$; $\sigma = 5 \quad \nu_1 = 250 \quad k_1 = 0.5$

Model	AIC	ν_1	ν_2	k_1	RMSE	σ	α
FT Skew Asy.	-342.97	230.21	2.25	0.9	0.75	4.33	0.17
FT Asy.	-327.04	225.32	1.43	0.47	0.6	4.33	0.5
FT Skew	-306.46	0.99	0.99	2.17	1.26	3.71	0.43
FT	-305.26	0.86	0.86	2.07	1.27	3.66	0.5
Skew	-39.34	250	250	0.1	10.77	483.02	0.03
Gaussian	39.79	250	250	0.1	10.77	506.4	0.5

Table 4.6: Simulation results with true parameters set to $\boldsymbol{\alpha = 0.5} \quad \boldsymbol{\nu_2 = 3}$; $\sigma = 5 \quad \nu_1 = 250 \quad k_1 = 0.5$

Model	AIC	ν_1	ν_2	k_1	RMSE	σ	α
FT Skew Asy.	-442.72	153.5	38.48	0.8	0.45	4.01	0.42
FT Asy.	-439.23	150.49	23.02	0.72	0.38	4.18	0.5
FT Skew	-443.4	49.55	49.55	1.06	0.59	3.85	0.37
FT	-439.37	41.72	41.72	1.04	0.6	3.92	0.5
Skew	-437.99	250	250	0.29	0.52	5.2	0.22
Gaussian	-422.38	250	250	0.3	0.53	5.74	0.5

Table 4.7: Simulation results with true parameters set to $\boldsymbol{\alpha = 0.4} \quad \boldsymbol{\nu_2 = 3}$; $\sigma = 5 \quad \nu_1 = 250 \quad k_1 = 0.5$

Model	AIC	ν_1	ν_2	k_1	RMSE	σ	α
FT Skew Asy.	-438.26	206.28	14.12	0.81	0.47	4.02	0.28
FT Asy.	-429.95	199.63	2.38	0.67	0.35	4.15	0.5
FT Skew	-437.39	22.88	22.88	1.26	0.65	3.63	0.26
FT	-426.02	6.01	6.01	1.11	0.63	3.87	0.5
Skew	-430.74	250	250	0.25	0.72	5.41	0.08
Gaussian	-401.75	250	250	0.26	0.73	6.39	0.5

Table 4.8: Simulation results with true parameters set to $\boldsymbol{\alpha = 0.3 \quad \nu_2 = 3}$; $\sigma = 5 \quad \nu_1 = 250 \quad k_1 = 0.5$

Model	AIC	ν_1	ν_2	k_1	RMSE	σ	α
FT Skew Asy.	-436.09	219.83	8.96	0.82	0.53	3.95	0.15
FT Asy.	-419.19	215.88	1.63	0.64	0.39	4.04	0.5
FT Skew	-427.77	12.6	12.6	1.53	0.71	3.38	0.28
FT	-413.76	9.56	9.56	1.36	0.75	3.6	0.5
Skew	-414.27	250	250	0.19	1.02	5.88	0.03
Gaussian	-370.98	250	250	0.2	1.04	7.54	0.5

Table 4.9: Simulation results with true parameters set to $\boldsymbol{\alpha = 0.5 \quad \nu_2 = 250}$; $\sigma = 5 \quad \nu_1 = 250 \quad k_1 = 0.5$

Model	AIC	ν_1	ν_2	k_1	RMSE	σ	α
FT Skew Asy.	-466.59	147.93	162.4	0.77	0.23	4.08	0.47
FT Asy.	-462.55	127.95	134.07	0.75	0.24	4.25	0.5
FT Skew	-467.09	173.79	173.79	0.74	0.2	4.2	0.47
FT	-463.8	143.6	143.6	0.73	0.2	4.31	0.5
Skew	-467.49	250	250	0.53	0.07	4.45	0.48
Gaussian	-464.1	250	250	0.52	0.06	4.58	0.5

Table 4.10: Simulation results with true parameters set to $\boldsymbol{\alpha = 0.4 \quad \nu_2 = 250}$; $\sigma = 5 \quad \nu_1 = 250 \quad k_1 = 0.5$

Model	AIC	ν_1	ν_2	k_1	RMSE	σ	α
FT Skew Asy.	-471.89	193.63	180.92	0.66	0.18	4.07	0.24
FT Asy.	-459.99	178.46	116.8	0.67	0.21	4.35	0.5
FT Skew	-472.04	189.51	189.51	0.69	0.17	4.12	0.27
FT	-461.24	145.06	145.06	0.65	0.18	4.41	0.5
Skew	-474.47	250	250	0.52	0.06	4.3	0.23
Gaussian	-462.05	250	250	0.5	0.06	4.63	0.5

Table 4.11: Simulation results with true parameters set to $\boldsymbol{\alpha = 0.3 \quad \nu_2 = 250}$; $\sigma = 5 \quad \nu_1 = 250 \quad k_1 = 0.5$

Model	AIC	ν_1	ν_2	k_1	RMSE	σ	α
FT Skew Asy.	-469.5	208	162.17	0.69	0.23	4.07	0.17
FT Asy.	-454.32	204.48	95.99	0.69	0.26	4.41	0.5
FT Skew	-470.87	182.87	182.87	0.73	0.22	4.11	0.17
FT	-454.88	133.44	133.44	0.71	0.22	4.48	0.5
Skew	-473.94	250	250	0.5	0.06	4.31	0.13
Gaussian	-455.11	250	250	0.49	0.06	4.8	0.5

new set of simulations examines the performance of the score driven approach with this asymmetric distribution compared to the classic approach of discarding outliers through t-tests.

The score driven approach yields a clear benefit when one tail exhibits a very low rate of decay

In this section a score driven model with an AST distribution featuring distinct tail parameters is applied to clean the VAT data using the smoothing algorithm discussed above. This nonlinear approach is compared to the more common strategy of discarding outliers through t-tests after estimating a Gaussian state space model. For this experiment, the synthetic data are generated using the Kalman filter where the parameters of the model are chosen to be close to the parameters estimated with the VAT data. In a second step, noise following an asymmetric Student's t-distribution is added. Different specification for the asymmetric Student's t-distribution are compared.

Table 4.12 shows the results. The score driven approach performs significantly better when the tail parameters are low, while its relative advantage decreases when the tail parameters get close to 3. Experimentation with the VAT data shows that its tail parameters are close to one for many industries, suggesting that the score driven approach should be better strategy here.

4.4.4 Averaging both cleaning approaches

Although measurement errors decrease as more firms submit their VAT returns, the final estimate remains very noisy. Consequently, there is no benchmark series which can be used to compare the cleaned series with, complicating the evaluation of the cleaning methods. This is in contrast to the Monte Carlo experiments of the previous section where the data generating process is know and hence the evaluation straightforward.

Table 4.12: Simulations investigating the efficiency of the cleaning method. Root mean square log approximation error (×100). The log approximation error is given by the difference between the log observation y_t minus the log smoothed observation $\hat{y}_t$. With the score driven approach the pseudo observation is generated with the iterative score driven smoothing scheme presented in the previous section; when observations are discarded with t-tests the pseudo observation is the signal in that period given by the Kalman smoother. Results from 100 simulations. Bold figures indicate the lowest value.

ν_2	α	Score driven approach	Discarding outlying observations
1	0.3	**4.48**	5.38
1	0.4	**4.49**	5.31
1	0.5	**6.38**	6.51
3	0.3	2.35	**2.34**
3	0.4	**2.07**	2.07
3	0.5	**1.9**	1.91
250	0.3	**2.1**	2.1
250	0.4	**1.87**	1.88
250	0.5	**1.82**	1.83

These experiments are useful for deciding which degree of flexibility in the distribution of the measurement error is reasonable, and how the approach compares to the standard strategy of using t-tests under a specific environment. But they should be used with care when drawing conclusions on the modelling approach most adapted to the VAT data because the nature of the measurement errors affecting them remains mostly unknown. In this context, where there is an important uncertainty regarding the appropriate modelling framework, a popular solution is to average estimates from a set of models instead of choosing a particular model assumed to be optimal.

The simplest approach to model averaging consists of using equal weights for all models, but weighting models using Bayesian Model Averaging (BMA) techniques typically works better, especially in an uncertain and changing environment. BMA techniques rely on models' likelihood functions, that is the likelihood of observing the data given a particular model (which includes regressors, relationship between variables and prior distributions). As such models are weighted depending on the likelihood that the observed data arise from their processes; hence, importantly, the data used for model evaluation must be common to all models.

The model selection problem faced in this chapter diverges from those typically encountered in the economic and forecasting literature because the uncertainty arises from the cleaning method, and the very nature of cleaning consists of altering the data. While BMA is used to overcome the uncertainty arising from model specification when there is no or very little uncertainty regarding the data, the uncertainty here arises from the data used for estimation and forecasting. It is difficult to know which dataset - the data when outliers are excluded using t-tests or the data arising from the score driven cleaning method - is closer to reality. In this context BMA is not applicable.

Given the high uncertainty surrounding the data and the inability to compare the different cleaning strategy to a common benchmark, since the true data are never observed, using equal weights in the averaging scheme is the only reasonable option.

4.4.5 Illustration of the cleaning strategies using one industry as a case study

This section illustrates both cleaning strategies using one industry as a case study. This industry is one of the largest in terms of gross value added for small and medium size firms.

T-test approach

The t-test approach is illustrated first. In this strategy outlying observations are discarded sequentially using t-tests. Outlying observations are detected by testing the standardised observation errors given by (4.11) for significance with a critical value of +- 3.3. The standardised observation errors resulting from the first round of estimation using the original VAT data are shown in the first panel of figure 4.3. Each of those points may be interpreted as two-tailed t-tests and the horizontal lines indicate the threshold of +/- 3.3. Three observations are outside this threshold and therefore considered as outliers.

A complimentary standard robustness check with Gaussian state space models consists of analysing the one-step ahead forecast errors given by

$$v_t^s = v_t/\sqrt{F_t}, \tag{4.18}$$

with F_t the variance of the prediction error.

The Kalman filter relies on normally distributed one-step ahead forecast errors, an assumption which can be verified by testing the standardised prediction errors given by (4.18) for normality using the skewness and kurtosis statistics. Bowman and Shenton (1975) show that if the normality assumption is respected these statistics should asymptotically be distributed as

$$\mathrm{S} \sim N(0, 6/n), \quad \mathrm{K} \sim N(0, 24/n),$$

where n is the sample size.

The second panel of figure 4.3 shows the standardised prediction error resulting from the first round of estimation. Panel three shows the histogram and kernel density estimate of these errors, while panel four shows their associated Q-Q plot. From these figures it is clear that the standardised prediction errors are not normally distributed. This is confirmed with skewness and kurtosis statistics of 3.33 and 15.56, which yield z-scores of 11.93 and 22.49, both clearly rejecting the null hypothesis of normally distributed errors.

In the second round of estimation the observations previously detected as outliers in the first round are discarded and replaced with missing values. The results are presented in figure 4.4. There is one outliers in the standardised observation errors, and the standardised prediction errors fail again to be normally distributed with skewness and kurtosis statistics of 5.2 and 14.

Discarding the outlier found in the second round of estimation and re-estimating the model yields the results presented in figure 4.5. No outliers are detected, and with skewness and kurtosis statistics of -0.06 and 2.62 (giving z-scores of -0.19 and -0.67) the null hypothesis that the standardised prediction error is normally distributed is not rejected any more. Since no new outliers are detected, the data after the second round of estimation can be used for estimating the nowcasting model.

Score driven approach

Score driven models provide an alternative approach for cleaning where outlying observations are downweighted instead of being discarded completely.

The estimated degrees of freedom with the score driven method are 250 for the left tail and 0.62 for the right tail. The degrees of freedom for the right tail is very low, but this is necessary to capture the extremely large outlying observations observable in the first plot. The left tail, on the other hand, is Gaussian, hence negative prediction errors are never downweighted and the pseudo observations are equal or below the original figures, but never above.

Figure 4.6 illustrates the downweighting behaviour of the score driven cleaning approach. The red line shows the downweighting applied to the data, which are shown using the vertical bars. Large outliers are downweighted whereas relatively small errors are not affected. Importantly negative prediction errors are not subject to any downweighting due to the very large tail parameter on the left side of the distribution.

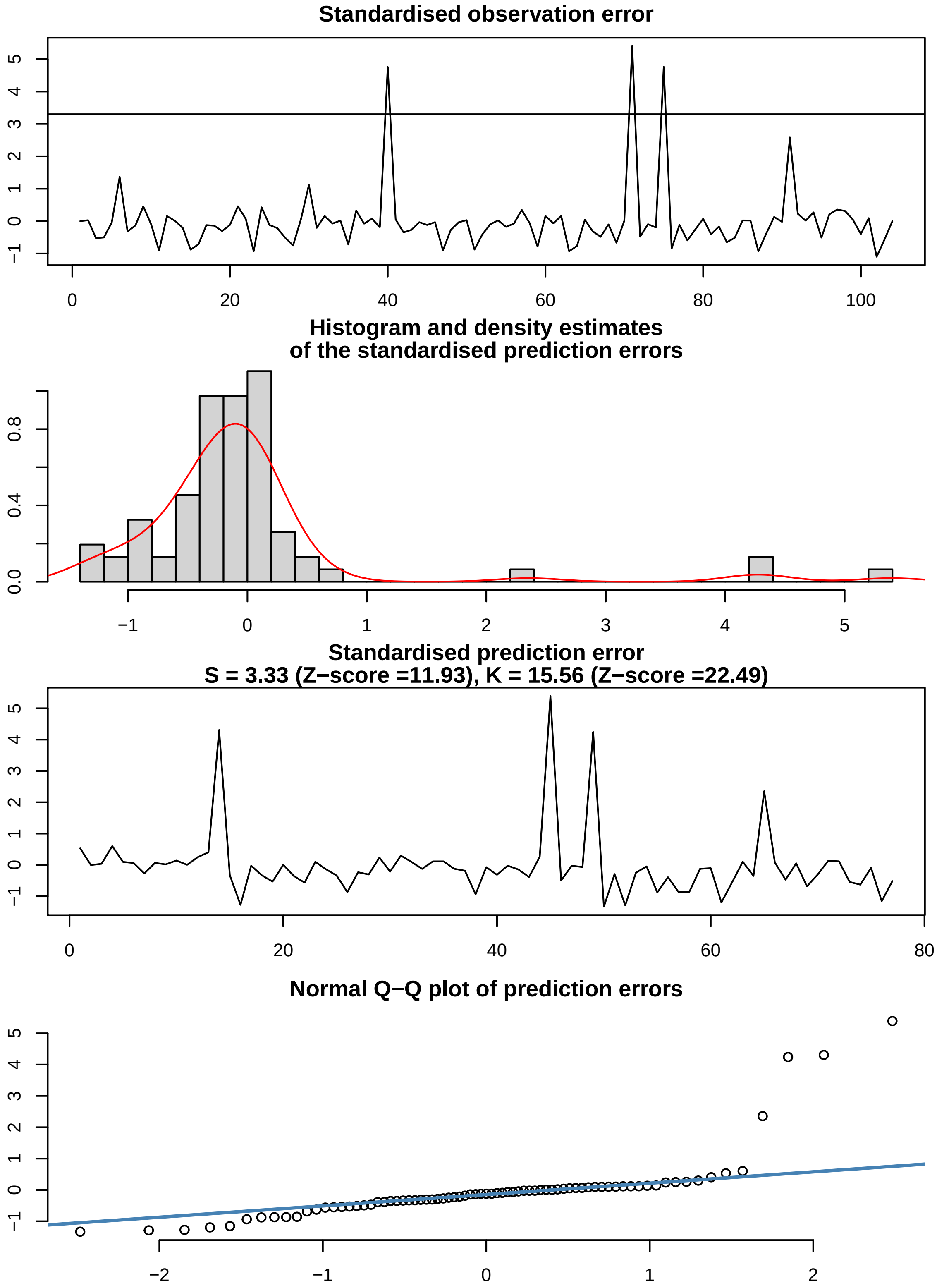


Figure 4.3: Results from the bivariate model estimated with VAT and MBS data for the transport industry. The VAT data are taken from the first release. The density of the standardised prediction errors is derived using a kernel density estimation with a Gaussian kernel (density() function in R). The first 25 standardised prediction errors are excluded because of the diffuse initialisation. Only outputs from the model with which it is not possible to re-construct the original data are shown to respect the confidentiality of the VAT respondents.

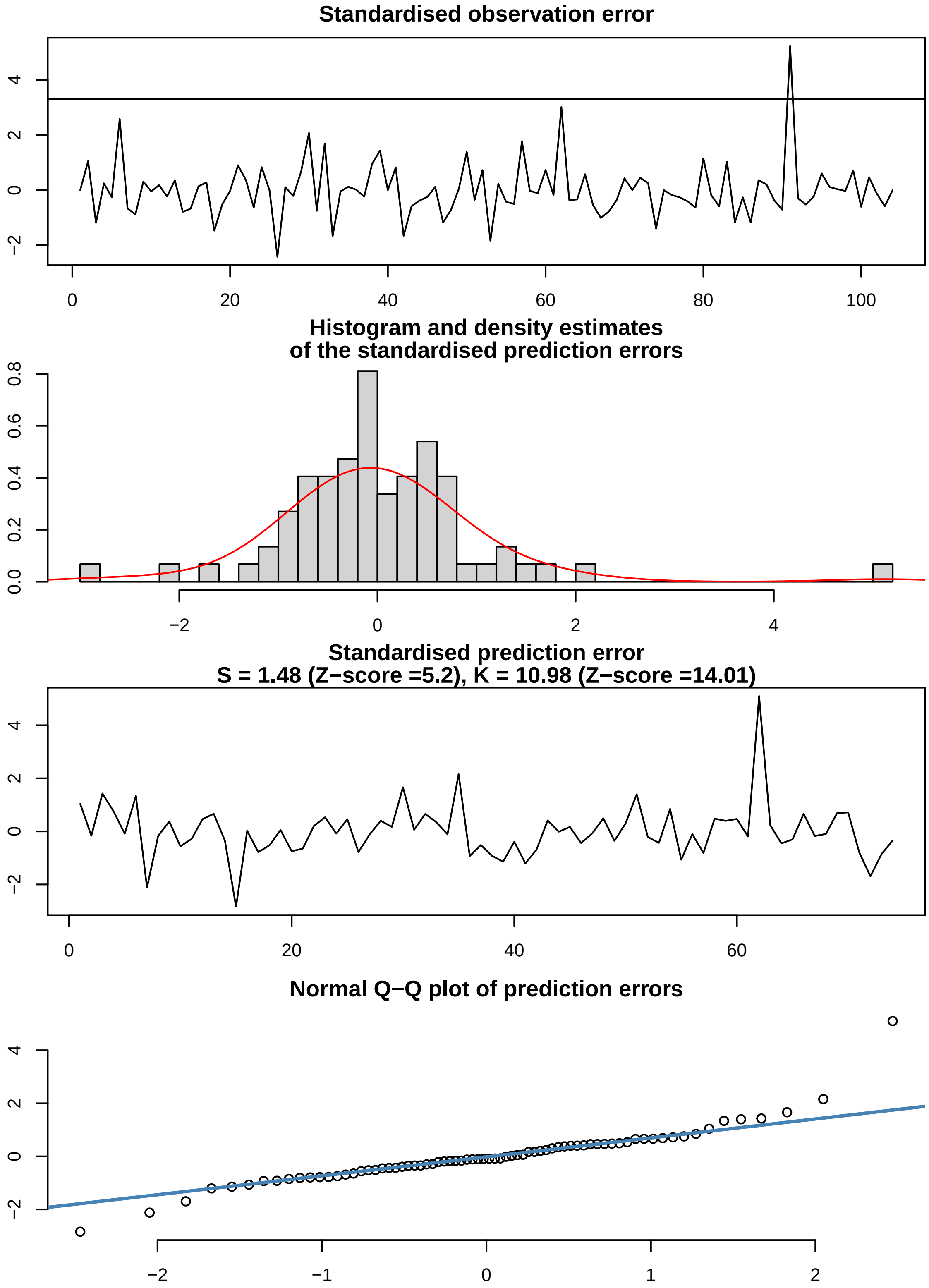
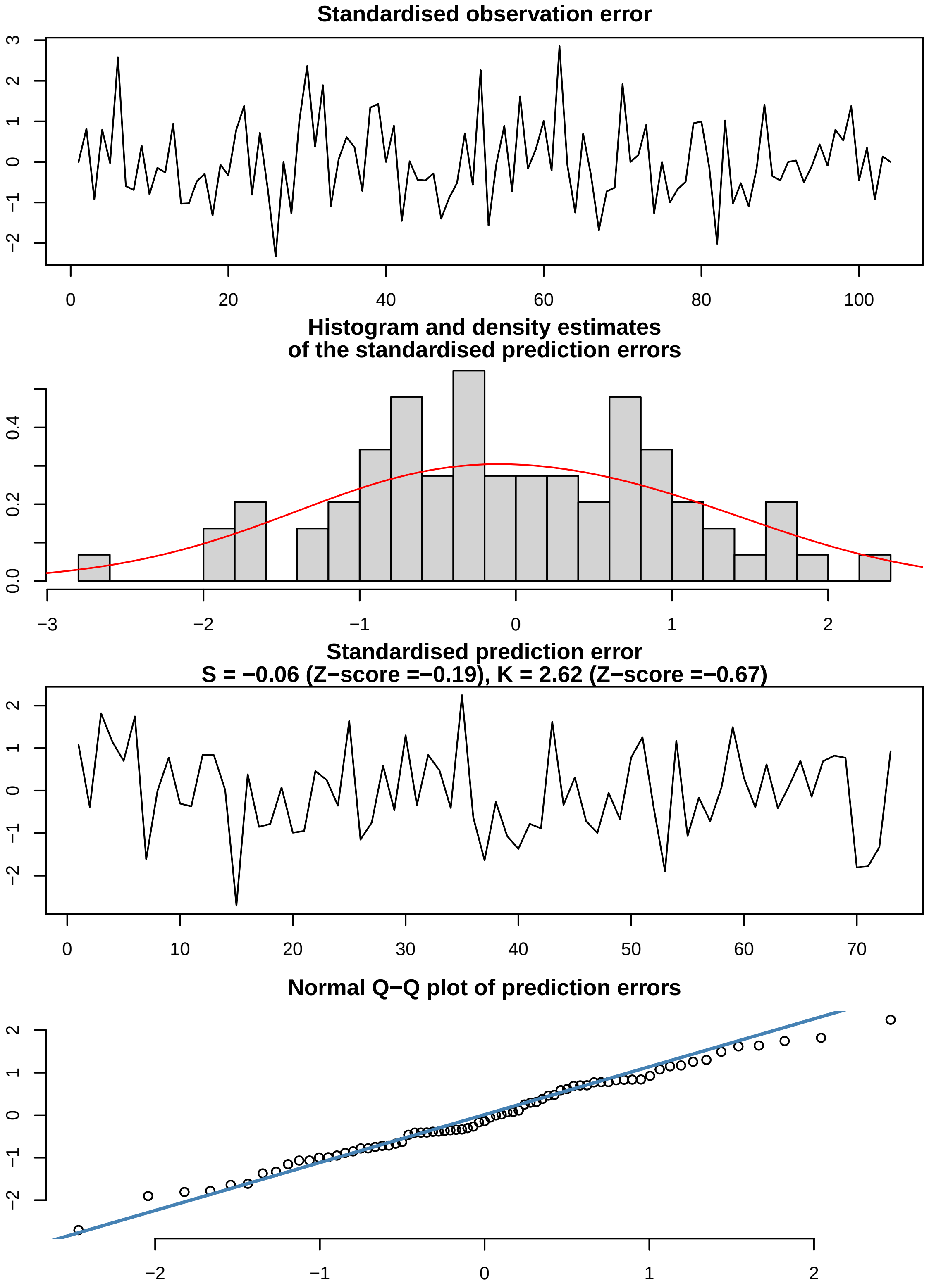


Figure 4.4: Results from the bivariate model estimated with VAT and MBS data for the transport industry. The VAT data are taken from the first release. **VAT figures at time points 40, 73 and 76 are discarded**. The density of the standardised prediction errors is derived with a kernel density estimation with a Gaussian kernel (density() function in R). Only outputs from the model with which it is not possible to re-construct the original data are shown to respect the confidentiality of the VAT respondents.


Standardised observation error

Histogram and density estimates
of the standardised prediction errors

Standardised prediction error
S = −0.06 (Z−score =−0.19), K = 2.62 (Z−score =−0.67)

Normal Q−Q plot of prediction errors

Figure 4.5: Results from the bivariate model estimated with VAT and MBS data for the transport industry. The VAT data are taken from the first release. **VAT figures at time points 40, 73, 76 and 93 are discarded**. The density of the standardised prediction errors is derived with a kernel density estimation with a Gaussian kernel (density() function in R). Only outputs from the model with which it is not possible to re-construct the original data are shown to respect the confidentiality of the VAT respondents.

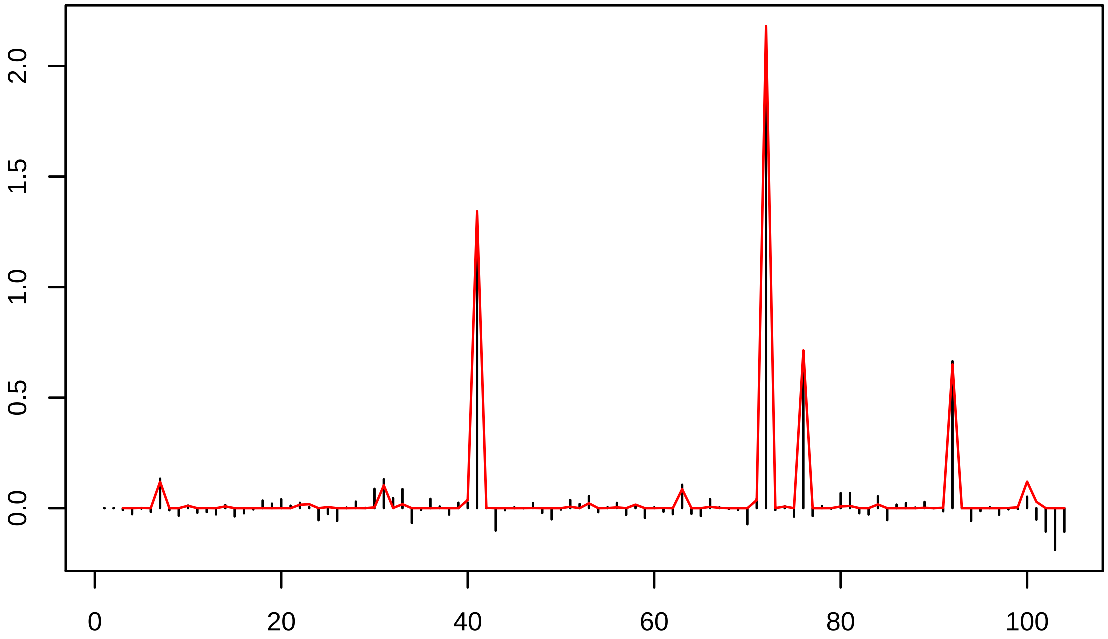


Figure 4.6: Illustration of the score driven cleaning method using VAT data from the first release (industry 493T495). The vertical bars show the prediction errors in log units derived from the score driven model with an asymmetric student distribution. The red line shows the difference in logs between the original VAT data and the pseudo ”clean” observations derived using the score driven smoothing algorithm. Only outputs from the model with which it is not possible to re-construct the original data are shown in order to respect the confidentiality of the VAT respondents.

4.5 Pseudo real-time exercise

The VAT data are regularly revised over time; i.e. the first estimate of the VAT observation for March 2020 using the first release of the quarterly March VAT observation (released in May) differs from the estimate when estimation is carried in July when the second and third release for May have been published, and quarterly estimates for April and May have become available. To investigate how the model would have performed in real time in May 2020 when forecasting monthly output for March 2020 it is important to use only the data that were available at the time. Proceeding this way it is then possible to analyse the revisions to the nowcasts as more data become available.

To analyse the performance of the model over time, estimation is carried out over a period of 40 months using 40 vintages of the data. Figure 4.7 illustrates two successive vintages. This is a *pseudo* real-time exercise because all VAT releases are cleaned together using the last vintages. Cleaning the VAT observations at each step of the rolling estimation in addition to estimating the nowcasting model would slow done the analysis excessively.

At each step of the rolling estimation the vector of seasonally adjusted monthly output estimates (both smoothed and filtered) in logs $\hat{x}^T = (\hat{x}_1^T, \hat{x}_2^T, ..., \hat{x}_N^T)'$, where T indicates the last month all maturities are observable and $N = T + 10$, are stored. By taking T forward gradually it is possible to simulate the real-time accrual of the data.

The revisions to the seasonally adjusted figures of interest are:

$$\text{revision}_{i,t}^{j} = \hat{x}_t^{11} - \hat{x}_t^{j}, \quad j = 1, ..., 10, \tag{4.19}$$

where $\hat{x}_t$ is the log of monthly seasonally adjusted output derived from the Kalman filter or smoother.

$$\begin{bmatrix} y_{11,1} & y_{10,1} & y_{9,1} & \cdots & y_{2,1} & y_{1,1} \\ y_{11,2} & y_{10,2} & y_{9,2} & \cdots & y_{2,2} & y_{1,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{11,t} & y_{10,t} & y_{9,t} & \cdots & y_{2,t} & y_{1,t} \\ {\color{red}y_{11,t+1}} & y_{10,t+1} & y_{9,t+1} & \cdots & y_{2,t+1} & y_{1,t+1} \\ & {\color{red}y_{10,t+2}} & y_{9,t+2} & \cdots & y_{2,t+2} & y_{1,t+2} \\ & & \ddots & \vdots & \vdots & \vdots \\ & & & {\color{red}y_{3,t+10}} & y_{2,t+10} & y_{1,t+10} \\ & & & & {\color{red}y_{2,t+11}} & y_{1,t+11} \\ & & & & & {\color{red}y_{1,t+12}} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_t \\ z_{t+1} \\ z_{t+2} \\ \vdots \\ z_{t+10} \\ z_{t+11} \\ {\color{red}z_{t+12}} \end{bmatrix}$$

Figure 4.7: Illustration of the successive vintages data used for estimation in a pseudo real-time setting. $y_{i,t}$ represents the quarterly VAT data from release i in month t. Columns and first subscripts indicate the releases while rows and second subscripts indicate the month the figures relate to. The maturity of the data increases from left to right and their timeliness increases from top to bottom.z_t represents the monthly covariate (the MBS for large businesses) in month t. In step t of the real-time pseudo exercise only black terms are available for estimation, while in step $t+1$ both black and red terms are used.

Two rounds of rolling estimation carried out, one for each data set (i.e. the one obtained when excluding outliers and the one obtained with the score driven cleaning approach). The revisions implied by both methods are analysed and compared with the revisions resulting from averaging nowcasts.

Sources of revision

There are four channels through which the accrual of releases can yield to revisions. First, noise in the aggregate figures should decrease with maturity, i.e. $y_{4,t}$ should be a better estimate of the signal $\tilde{y}_{1,t}$ than $y_{1,t}$ is. More precise aggregate figures should yield better estimates of the underlying seasonally adjusted figures.

Secondly, as revised aggregate figures for month t become available, early figures for periods succeeding t also become available. Notably, when all releases for month t are available, estimates of aggregate figures for periods up to $t+10$ are also available (as illustrated in figure 4.7). Assuming that the target is the seasonally adjusted monthly estimate for month t, the data succeeding t will affect the smoothed estimate of the target. This is because the smoothed estimate is the best estimate given past, current and future observations.

Another source of revision arising from using data succeeding the targeted period and affecting the smoothed estimate comes from the temporal aggregation constraints. All monthly figures are related through the temporal aggregation constraints because the observations are overlapping; modifying one quarterly figure affects the entire series of seasonally adjusted estimates. For these reasons large revisions to the smoothed series should not be taken to imply that the early releases are uninformative.

4.5.1 Aggregating then nowcasting

This section reports the results when cleaned VAT observations are aggregated across all seventy five industries before carrying out the real-time estimation exercise. This approach contrasts with nowcasting output for each industry and then aggregating these nowcasted figures.

The first plot of Figure 4.8 shows the original eleven releases of VAT observations taken from the most recent vintage. The figures represent all 75 industries and the data have been aggregated using gross value-added weights. Missing observations in the original series have been replaced with the average observation of the series for aggregation purpose. This plot illustrates well the extend and magnitude of the noise affecting the VAT data.

The second plot shows the same observations where extreme outliers have been discarded using t-tests, while the third plot shows the pseudo observations derived with the score driven method. From these two plots it is possible to get a much better picture of the VAT-based quarterly turnover.

The score driven cleaning approach produces smaller revisions and averaging does not help

Table 4.13 shows the mean absolute revision derived with (4.19) across each early release. These are revisions to the monthly seasonally adjusted output estimate in logs. The score driven method produces the lowest revisions in absolute values on average. Importantly the revisions' magnitude decreases monotonically as the maturity of the data increases, with a mean absolute revision of 1.3 percentage point for the first release, compared to 0.06 percentage point for the last.

Table 4.13: Mean absolute revision (×100) each cleaning approach and across releases. The revisions represent the differences in logs between the early estimates ($i = 1, ..., 11$) and last estimate ($i = 12$) of monthly seasonally adjusted output (equation (4.19)). Bold figures indicate the lowest mean absolute revision at each maturity.

Release	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$
Average	1.335	1.2	0.829	0.636	0.402	0.333	0.273	0.239	0.165	0.126
Score driven	**1.306**	**1.187**	**0.509**	**0.403**	**0.261**	**0.227**	**0.175**	**0.15**	**0.099**	**0.06**
T-test	1.692	1.532	1.315	1.006	0.626	0.508	0.42	0.37	0.278	0.213

Early VAT-based estimates tend to be revised towards

Table 4.14 shows the mean revision with its standard error and associated t-statistic across each early releases. While the absolute revision is useful to analyse the magnitude of the revisions, the mean revision and its standard error are used to investigate the biases in the early releases. Here again the score driven method performs better than the other two approaches as it tends to produce lower means and standard errors.

But unfortunately the first five releases exhibit a significant negative bias. This is not necessarily a bad feature because it can arise from relatively small standard errors, indicating a low volatility in the mean revisions which is desirable since these are small. However, this also implies that the monthly output estimates tend to be revised downwards as more data become available.

Table 4.14: Mean revision and standard error (×100) with implied t-statistic for each cleaning approach and across releases. The revisions represent the differences in logs between the early estimates ($i = 1, ..., 11$) and last estimate ($i = 12$) of monthly seasonally adjusted output (equation (4.19)). Bold figures indicate the lowest mean revision at each maturity.

Cleaning	Stat.	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$	$j=10$
Average	Mean	-0.931	-0.906	-0.532	-0.336	-0.198	-0.155	-0.12	-0.115	-0.084	-0.067
Average	S.E.	0.274	0.214	0.16	0.12	0.067	0.059	0.053	0.046	0.036	0.032
Average	t-stat	-3.397	-4.227	-3.327	-2.795	-2.982	-2.616	-2.249	-2.522	-2.319	-2.082
Score driven	Mean	-0.977	**-0.79**	**-0.286**	**-0.175**	**-0.123**	**-0.075**	**-0.046**	**-0.033**	**-0.029**	**-0.021**
Score driven	S.E.	0.283	0.229	0.085	0.07	0.049	0.045	0.038	0.034	0.02	0.014
Score driven	t-stat	-3.454	-3.454	-3.359	-2.503	-2.484	-1.659	-1.215	-0.96	-1.42	-1.483
T-test	Mean	**-0.875**	-1.015	-0.773	-0.495	-0.275	-0.234	-0.193	-0.197	-0.138	-0.113
T-test	S.E.	0.336	0.294	0.272	0.202	0.113	0.101	0.09	0.081	0.071	0.063
T-test	t-stat	-2.609	-3.455	-2.846	-2.455	-2.443	-2.314	-2.137	-2.428	-1.949	-1.802

Figure 4.9 helps to shade light on the nature of this negative bias; it shows the first release of monthly output alongside the last release and the implied revision. While both estimates tend to be close, there are few discrepancies where the first release shows a large increase in output which does not materialise in the last release. This is probably due to the cleaning approach which fails to capture large measurement errors in these few months. The negative bias is likely to be driven by these few large discrepancies. These large negative revisions can also be seen in figure 4.10 which shows output estimates from each releases across the entire real-time estimation period.

The VAT data can provide a timely indication of economic slumps

Although the VAT-based monthly output estimates tend to be revised towards, they can still provide a timely picture of economic activity. Figure 4.11 shows the first releases of the VAT-based monthly output estimates alongside the estimate arising from the MBS. The VAT-based estimate indicate a clear important reduction of economic activity starting from March 2020, albeit on of a smaller magnitude than the picture given by the MBS.

Overtime the VAT and MBS-based monthly output signals coincide but there remains some differences in their levels

Figure 4.12 shows the monthly output figures derived using the latest available vintage of the VAT data, that is in the last step of the real-time estimation exercise. Here both cleaning strategies are illustrated. These series are compared with the monthly output estimate derived with the MBS data. The first plot shows the levels while the second plot shows log differences.

Both VAT-based series show similar trends, but these diverge locally with the trend underlying the MBS figures. This is consistent with the result obtained in chapter 3. Separately, the score driven approach produces monthly changes closer to the MBS monthly changes, with a correlation coefficient of 0.87 when using the score driven approach compared to 0.76 when using the t-test cleaning. Hence, in addition to producing lower revisions, the score driven approach yields a picture of economic activity closer to the MBS that the t-test cleaning method.

The covariate helps in estimating VAT-based monthly changes

Table 4.15 shows the estimation results when estimating the nowcasting model (4.9) with VAT observations cleaned using the score driven method, since the previous section shows that it is the cleaning method which yields the lowest revisions and therefore is preferable.

The first row shows variance and correlation estimates of the local linear trend model, which estimated smoothed components are plotted in figure 4.13. The VAT and MBS disturbances in the local linear trend model governing the monthly seasonal adjusted output estimates are strongly correlated. This highlights the importance of the MBS as covariate for improving VAT-based monthly output.

The bias in the first release changes rapidly

The second row of table 4.15 shows the variance estimates of the releases biases' disturbances, while the release biases are plotted in figure 4.14. These show that the bias in the first release changes significantly over time, a phenomenon less pronounced in more mature releases.

The first two releases are significantly more noisy than later ones

The third row shows the estimated variance of the releases' observation errors. The first and second releases are much more noisy than later releases. Separately, estimated noise remains relatively strong in the last release, which shows that even mature figures are subject to non-negligible measurement errors.

Table 4.15: Estimation results from the nowcasting model using the latest vintage of VAT data cleaned with the score driven approach. The first line shows the estimated variance and correlation parameters of the bivariate local linear trend model at the core of the nowcasting model. The second line shows the estimated variance parameters of the disturbances in the release biases. The third line shows the estimated variance parameters of the release observation errors. Variance parameters are reported $\times 1e4$.

$\sigma_{\mu,1}$	$\sigma_{\nu,1}$	$\sigma_{e,1}$	$\sigma_{\mu,2}$	$\sigma_{\gamma,2}$	$\sigma_{\nu,2}$	$\sigma_{e,2}$	ρ_{ν}	ρ_{μ}	ρ_{e}	
0.620	0.090	7.990	3.970	0	0.110	5.220	1	1	0.930	
$\sigma_{c,1}$	$\sigma_{c,2}$	$\sigma_{c,3}$	$\sigma_{c,4}$	$\sigma_{c,5}$	$\sigma_{c,6}$	$\sigma_{c,7}$	$\sigma_{c,8}$	$\sigma_{c,9}$	$\sigma_{c,10}$	
0.130	0.030	0.010	0.010	0.010	0	0	0	0.010	0.060	
$\sigma_{\epsilon,1}$	$\sigma_{\epsilon,2}$	$\sigma_{\epsilon,3}$	$\sigma_{\epsilon,4}$	$\sigma_{\epsilon,5}$	$\sigma_{\epsilon,6}$	$\sigma_{\epsilon,7}$	$\sigma_{\epsilon,8}$	$\sigma_{\epsilon,9}$	$\sigma_{\epsilon,10}$	$\sigma_{\epsilon,11}$
5.620	5.860	0.170	0.270	0.040	0.080	0.210	0.370	0.380	0.090	0.210

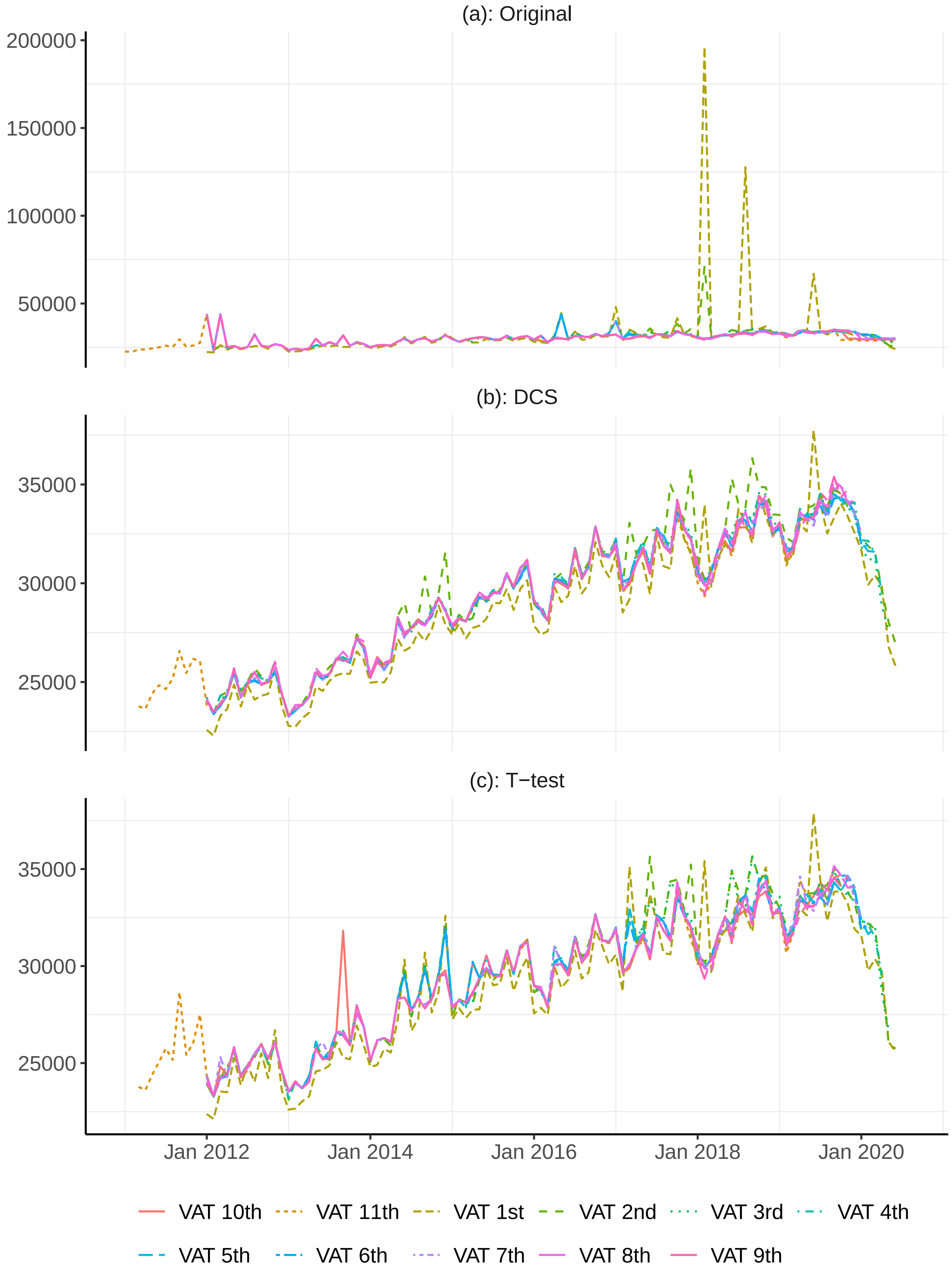


Figure 4.8: Aggregate non-seasonally adjusted VAT releases (representing approximately a quarter of GVA in the UK). Original VAT observations (most recent data vintage) compared with the observations obtained when using the t-test and score driven cleaning approaches.

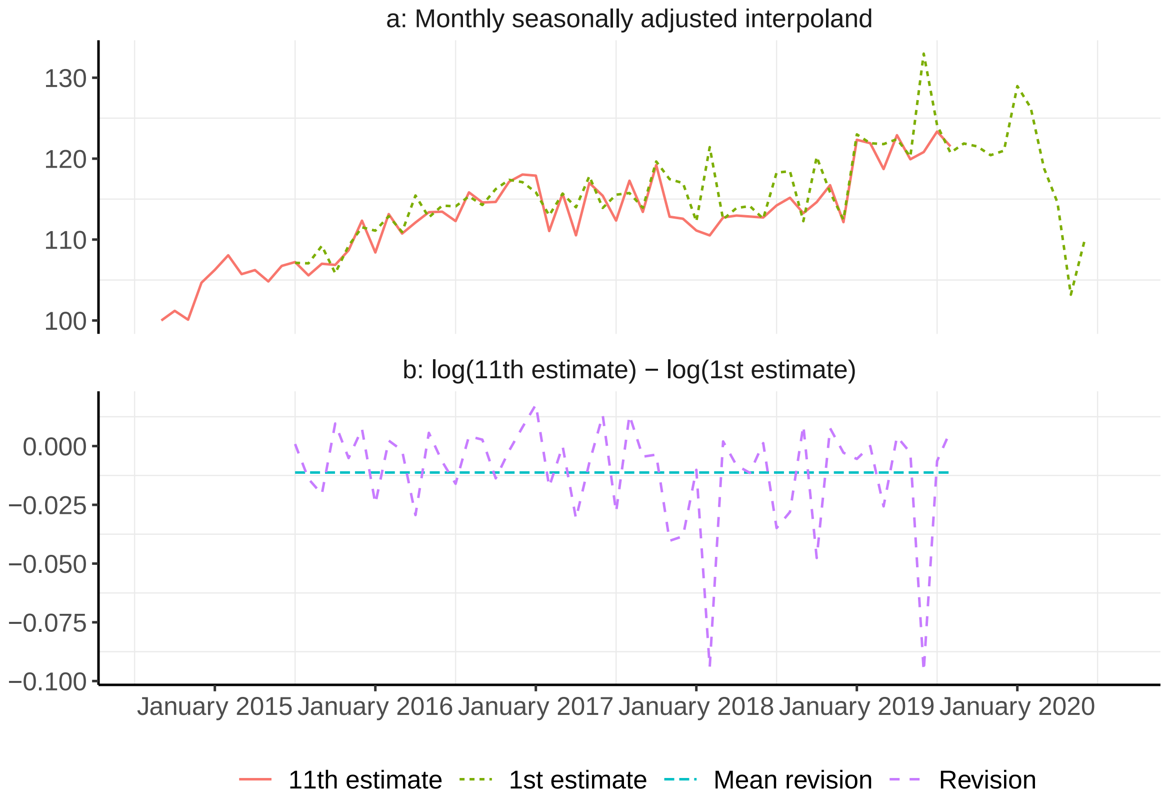


Figure 4.9: Seasonally adjusted figures representing small and medium size businesses in 75 industries (a quarter of GVA in the UK). First and last (eleventh) VAT estimates. Estimation using pseudo real-time data. The VAT-based figures are derived using the score driven method for cleaning. Index July 2014=100.

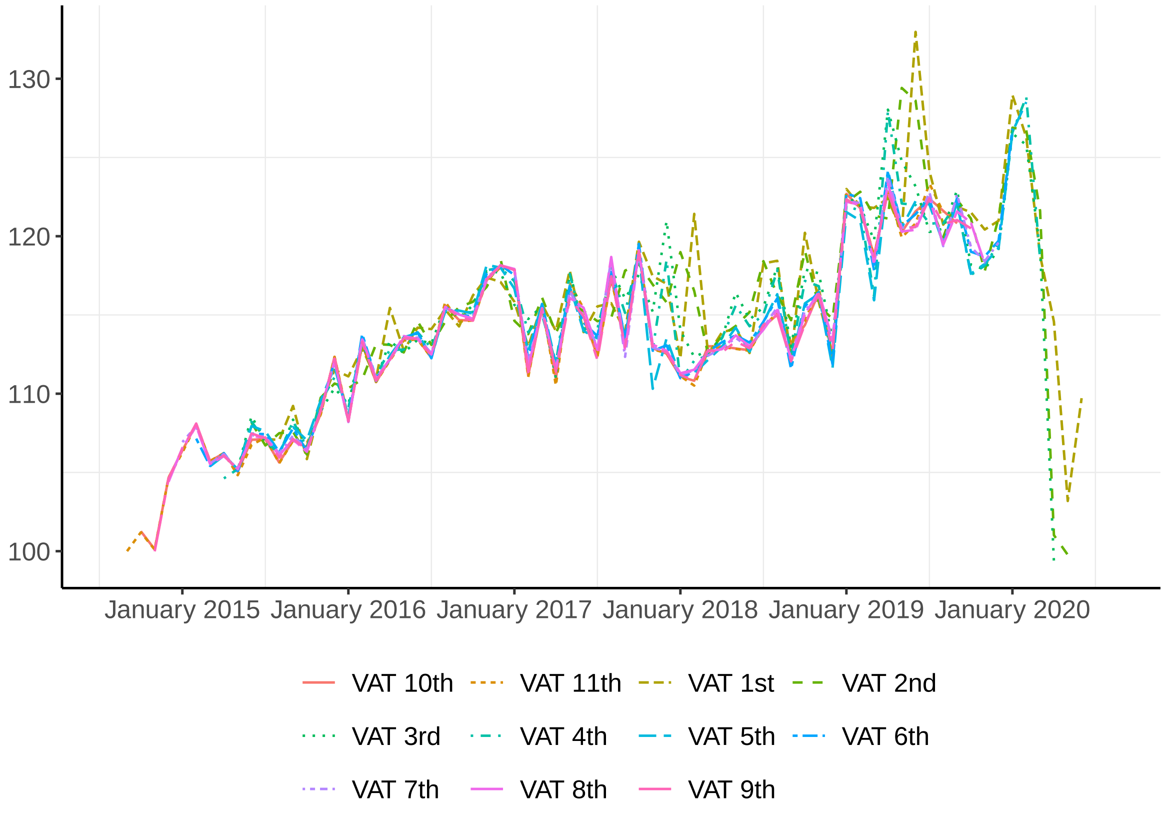


Figure 4.10: VAT-based nowcasted monthly output. Seasonally adjusted figures representing small and medium size businesses in 75 industries (a quarter of GVA in the UK). First VAT estimate with the next ten revisions. Estimation using pseudo real-time data. The VAT-based figures are derived using the score driven method for cleaning. Index July 2014=100.

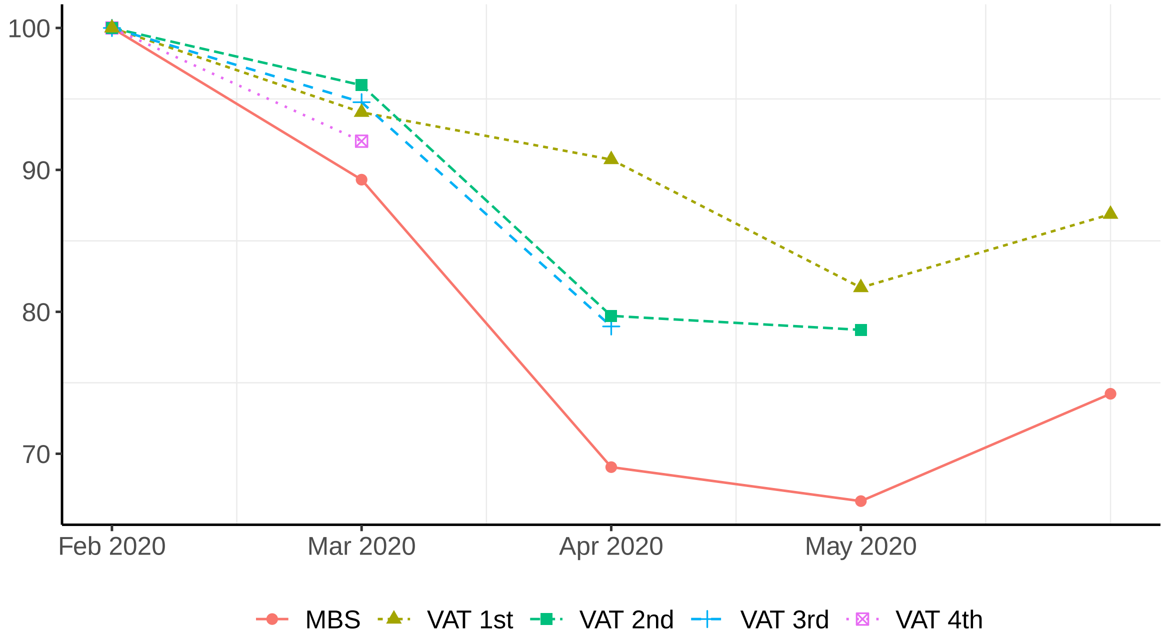


Figure 4.11: Real-time nowcasts of monthly output. Seasonally adjusted figures representing small and medium size businesses in 75 industries (a quarter of GVA in the UK). First VAT estimate and the four subsequent revisions. The VAT-based figures are derived using the score driven method for cleaning. Index February 2020=100.

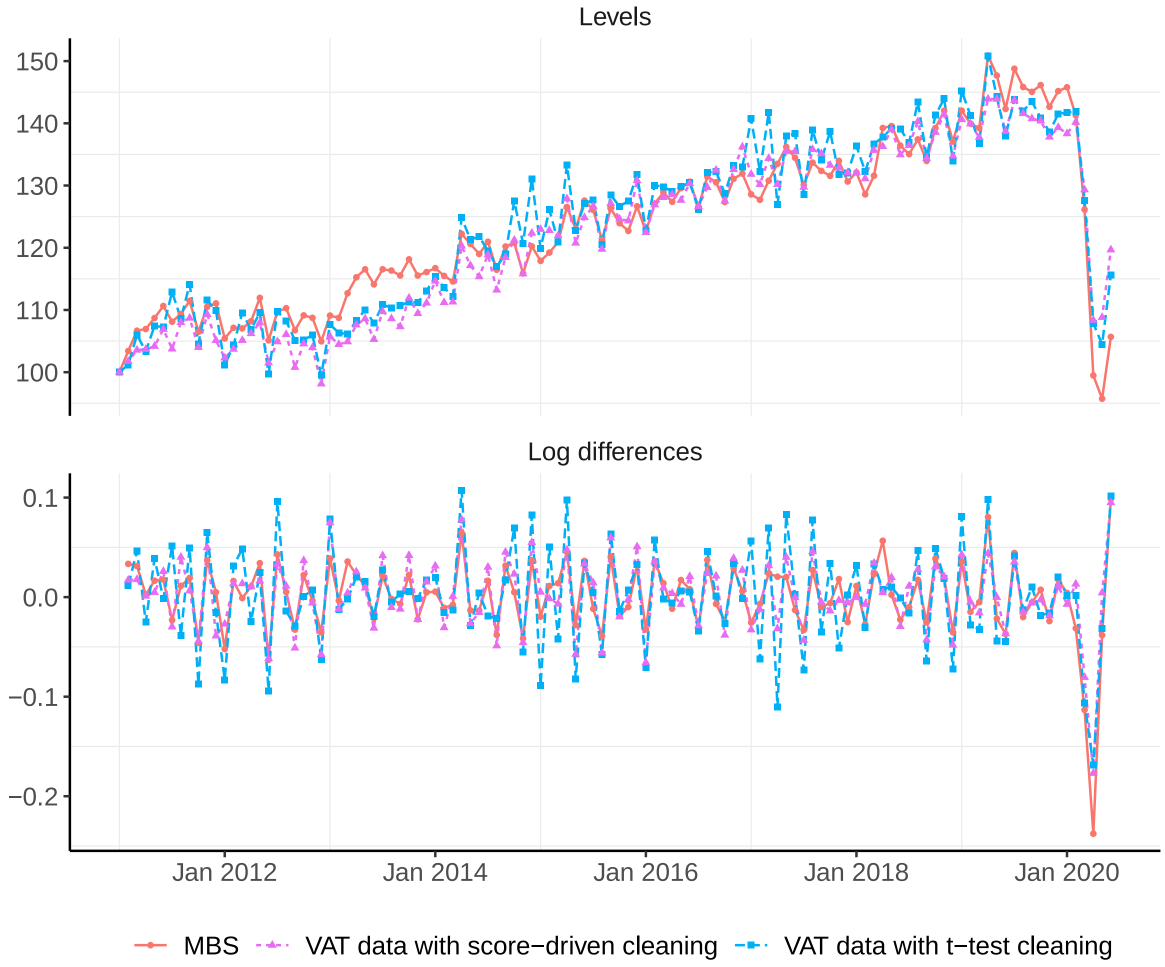


Figure 4.12: VAT-based and MBS-based nowcasted monthly output. Seasonally adjusted figures representing small and medium size businesses in 75 industries (a quarter of GVA in the UK). The VAT data are taken from the most recent vintage. Index January 2011=100. Corr log differences MBS and VAT when score driven cleaning = 0.87; Corr log differences MBS and VAT when t-test cleaning = 0.76.

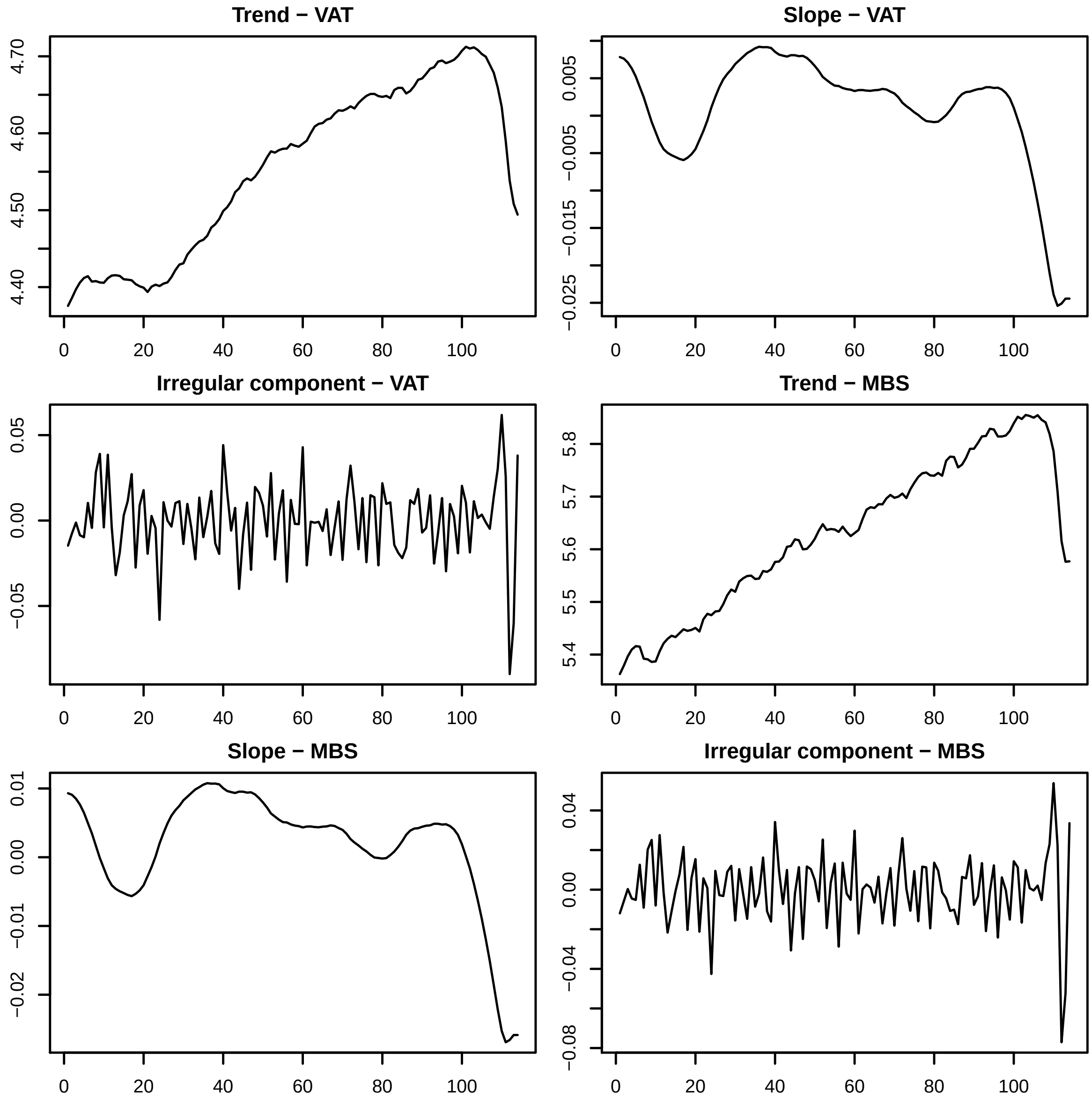


Figure 4.13: Smoothed estimates, in log units, of the bivariate local linear trend model at the core of the nowcasting model. Estimation with the most recent data vintages which have been cleaned using the score driven method and aggregated across all 75 industries using gross value-added weights. Monthly seasonally adjusted figures are derived by adding up the trend with the irregular component.

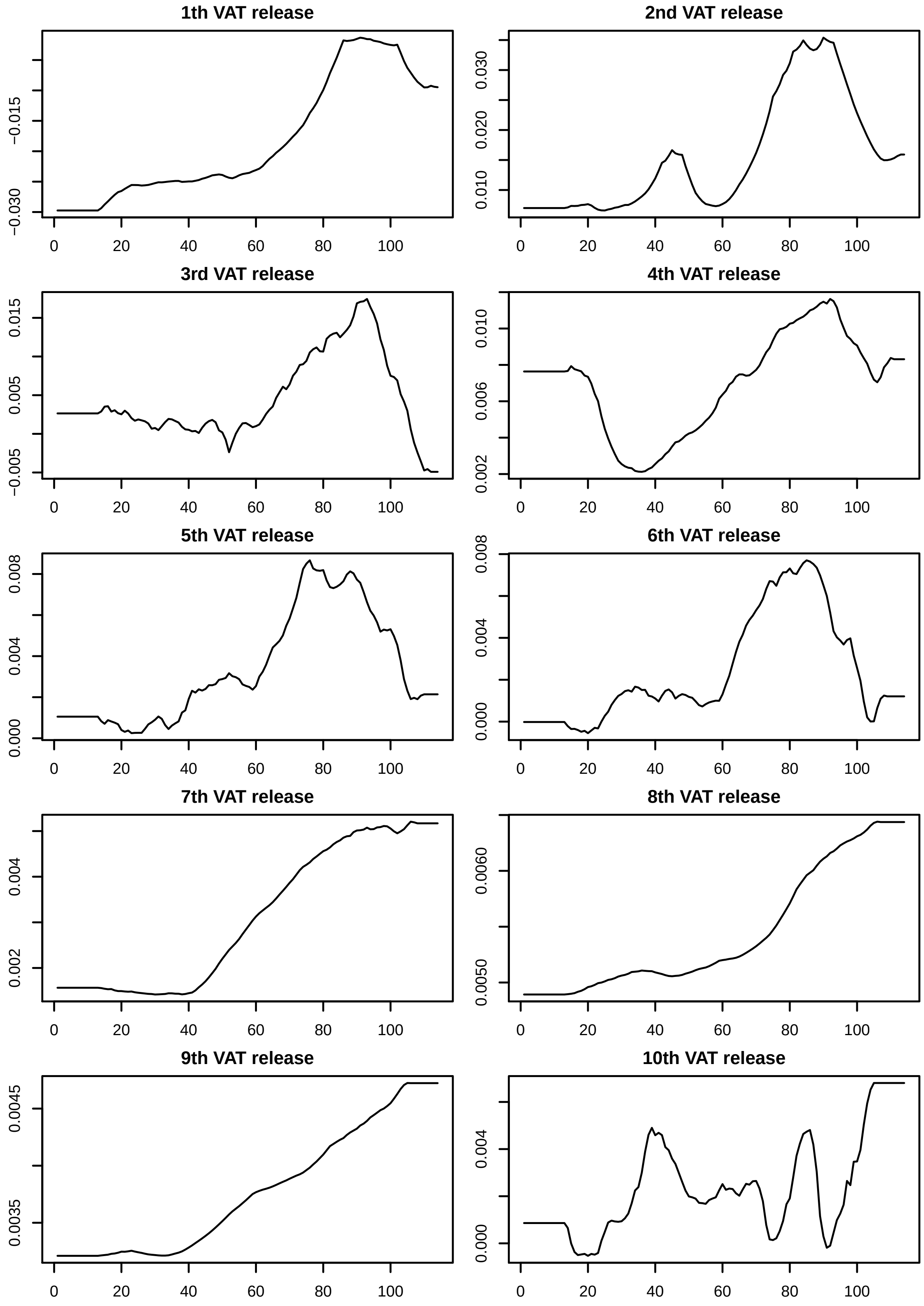


Figure 4.14: Bias of the first ten VAT releases with respect to the eleventh (and last) release. Log units. Estimation with the most recent data vintages which have been cleaned using the score driven method and aggregated across all 75 industries using gross value-added weights.

4.6 Conclusion

This chapter has presented and illustrated a flexible nowcasting approach for data subject to large measurement errors and revisions. It has shown that, in the presence of extremely large and asymmetric measurement errors, using a score driven approach works better than discarding outlying observations. This finding is illustrated through the production of a monthly output series from noisy VAT-based quarter figures which can be used as a timely indicator of the economic recession resulting from the coronavirus pandemic.

Chapter 5

Capturing GDP nowcast uncertainty in real time

5.1 Introduction

This chapter proposes a novel means for capturing changes in the shape of GDP nowcast uncertainty in a timely way by exploiting forecasting errors observed in series related to economic growth. The main methodological contribution lies in setting out a strategy for modelling relationships in scale and shape parameters, controlling the dispersion and asymmetry of density forecasts, across potentially heterogeneous series sampled on different frequencies and released asynchronously. Nowcasts can thus be communicated with an improved appreciation of the statistical uncertainty attached to them. The score driven model presented in this chapter is able to extract a timely indication from employment data of the asymmetry and dispersion attached to US GDP nowcast uncertainty during the onset of the coronavirus pandemic.

GDP is the most comprehensive measure of economic activity. As such it is a critical variable helping policy makers - and economic agents in general - make decisions. But the quality of GDP comes at the cost of timeliness; it is published with a significant delay. Thus economic agents typically make use of forecasting methods to get an idea of the GDP number before its publication by the national statistics office. When the number of interest refers to the current quarter or recent past, forecasting is referred to as 'nowcasting'. Nowcasting methods rely on series related to economic growth but released earlier and possibly more frequently than GDP, such as employment data.

There are two widely used approaches for nowcasting: Dynamic factor models and MIDAS regressions. Dynamic factor models use a small number of common factors to capture the relationship between related series and GDP. New observations on the related series lead to updates in the common factors which in turn lead to updates in GDP nowcasts. To handle the mixed-frequency nature of the model, the low frequency variables are modelled endogenously at the highest frequency of observation in the data using unobserved components methods and linked to the low frequency observations through temporal aggregation constraints. On the one hand, MIDAS regressions use related series as independent variables which are aggregated to the frequency of the dependent variable using arbitrary lag polynomials.

For an efficient use of nowcasts, their associated uncertainty should be conveyed as well. Both methods discussed above provide distinct ways to use related series to improve not only GDP point nowcasts, but their associated uncertainty as well. In dynamic factor models this is done by introducing time-variation in the variance of the common factors, yielding stochastic volatility models (see notably Antolin-Diaz et al. (2017, 2020)). In MIDAS models, Pettenuzzo et al. (2016) show how stochastic volatility can be introduced by using a parallel MIDAS regression for the scale parameter of the dependent variable in addition to the one typically used for the location parameter.

While stochastic volatility models yield timely changes in the dispersion of density nowcasts, they ignore a third dimension of forecasting uncertainty: its occasional asymmetry. By conditioning on financial variables, Adrian et al. (2019) and Delle-

Monache et al. (2020) notably find increased negative skewness accompanying a rise in volatility in the predictive distribution of US GDP growth in times of recessions. In normal times, however, GDP growth is close to being conditionally normally distributed. The sharp downturn brought by the coronavirus pandemic and the increase in asymmetry of forecasting uncertainty linked to it makes this issue particularly salient.

This chapter shows how the asymmetry of GDP nowcast uncertainty can be captured by modelling a common factor in the shape parameters, which control the asymmetry of the predictive distributions, of both GDP and a timely related series. This is in addition to common factors in location and dispersion parameters. Hence new observations in the related series lead to timely updates in the point forecasts through the common factor in the locations and, importantly, to the uncertainty attached to them as well through the common factors in the scales and shapes. Figure 5.1 shows how this approach can be used to capture the uncertainty during the onset of the coronavirus pandemic. It was clear in April 2020 that first-quarter GDP growth in the US would be negative or very close to zero, but while the magnitude of the drop was uncertain, a positive surprise was clearly improbable. This forecasting environment is reflected by a skewed predictive distribution towards negative values shown on the left panel. This contrasts with the right panel where only the dispersion and location of the predictive distribution can vary.

To capture cross-sectional dependencies in the data and nowcast GDP this chapter builds on the score driven approach of Harvey (2013) and Creal et al. (2013). Thus it provides an alternative nowcasting route to dynamic factor models and MIDAS regressions where common dynamics in skewness has not yet been explored. Score driven models provide a flexibility close to non-Gaussian state space models while remaining easy to estimate and implement. They have been applied successfully to economic forecasting problems notably by Delle-Monache and Petrella (2017), Creal

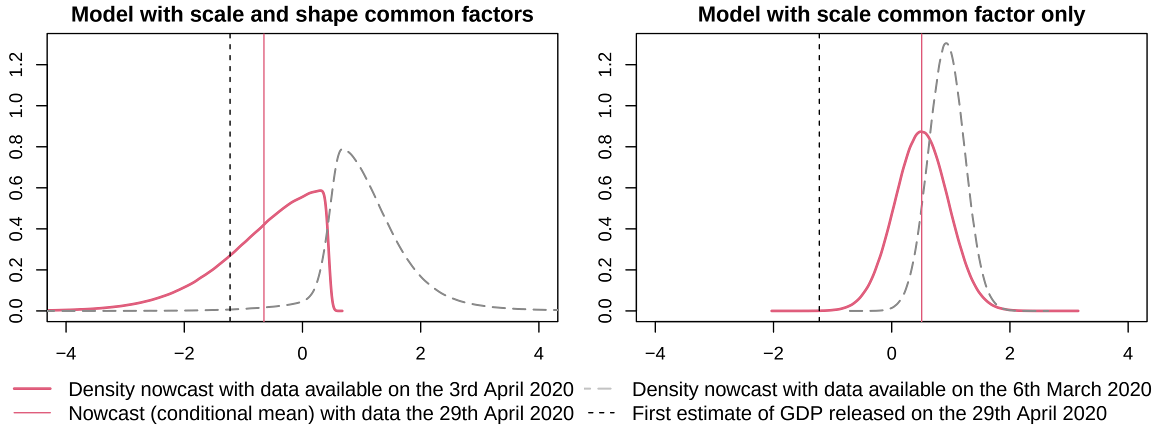


Figure 5.1: Density nowcasts for 2020 Q1 US GDP. The density in the left panel is generated by the Skew t model with stochastic volatility and shape common factor (model (a) below). The right panel is generated by the Student's t model with stochastic volatility (model (c))

et al. (2014), Gorgi et al. (2019) and Delle-Monache et al. (2020). However, their use for nowcasting has so far been limited. Indeed, score driven models are predictive filters where dynamic parameters are perfectly predictable conditional on past information. This means that contemporaneous information on the related series (e.g. data on employment for March when nowcasting first-quarter GDP) cannot in theory be used to improve nowcasts. However, Buccheri et al. (2021) show that score driven models can be seen as approximate filters and propose to use a filtering equation analogous to the filtering equation of the Kalman filter to update current parameter estimates following the release of contemporaneous data. This is the strategy adopted here.

Although GDP growth and the related series used to improve nowcasting exhibit a high degree of comovement, they can have heterogeneous features. Accordingly Creal et al. (2014) propose a score driven dynamic factor model where each series can have a distinct conditional distribution. However, their approach relies on independent prediction errors across series. This assumption is problematic when using correlated macroeconomic series subject to common shocks such as the sharp fall in activity induced by the pandemic. To model cross-sectional dependencies in the prediction errors following different conditional distributions this chapter makes use of the idea presented in chapter 6 where a copula is used to capture conditional cross-dependence.

Separately, modelling common components in series aggregated and sampled at different frequencies raises temporal aggregation issues somewhat different for location, scale and shape parameters. While Mariano and Murasawa (2003) popularised a precise approximation for conditional means or location parameters, similar solutions for scale and shape parameters have not been discussed in the literature. Carriero et al. (2016, 2018, 2020) and Huber (2016) model common scale components but not in a mixed-frequency setting. While Gorgi et al. (2019) model a common scale component in a score driven bivariate model, they do so by constraining the scale parameters to be identical in both series and do not discuss the issue of temporal aggregation.

This chapter explores two different strategies to tackle the temporal aggregation problem of scale and shape parameters. First, it is shown how scale parameters may be aggregated temporally in Gaussian models using a convenient approximation, thus providing an approach for modelling volatility common factors in mixed-frequency dynamic factor models. The second approach consists of aggregating monthly series into rolling-quarterly figures. The mismatch in the frequency of aggregation is thus alleviated, and scale and shape common factors can be modelled in a non-Gaussian setting.

The mixed-frequency dynamic factor model presented in this chapter exploits two US time series: GDP and the index of total aggregate weekly hours. GDP is quarterly while the index of total aggregate weekly hours is monthly. All series are derived from real-time vintages provided by the Federal Reserve Bank of Philadelphia. Different specifications of the model are applied in a real-time setting to study the effect of modelling common factors in scale and shape parameters on nowcasting performance, with a particular interest in density forecasts and recessionary episodes.

Results show that modelling common factors in scale and shape parameters perform better than models with only location common factors towards the end of the nowcasting window, while providing competitive performances in other times. Scale and shape common factors prove to be particularly important to capture the forecasting uncertainty generated by the coronavirus pandemic. Finally, modelling fat tails in series related to economic growth can complicate the identification of turning points in activity in monthly models.

5.2 A Mixed-Measurement (Quasi) Score Driven Approach for Nowcasting with Location, Scale and Shape Common Factors

Common factors are used for exploiting dependencies across GDP growth and a related series in location, scale and shape parameters. A popular method for estimating unobserved components such as common factors consists of writing the model in state space form and using the Kalman filter to evaluate the model's log likelihood. The Kalman filter provides an efficient approach for estimating nowcasting models for two reasons. First, it handles easily the missing values that arise when modelling jointly series aggregated at different frequencies and released asynchronously. Secondly, predictions can be decomposed into latent components, which notably can be used to capture secular changes in addition to common factors (see Harvey (1989) and Durbin and Koopman (2012)). A comprehensive presentation of dynamic factor and state space methods for nowcasting is given by Banbura et al. (2013).

A limitation of the Kalman filter, however, is that it relies on the data being conditionally normally distributed (prediction errors from the model must be normally distributed). Non-Gaussian features can be introduced using importance sampling methods but these can be computationally intensive. Alternatively, Creal et al. (2013) and Harvey (2013) derive a new class of filters relying on the score of the predictive log likelihood function which can arise from a wide range of families. Score driven models provide a general framework for introducing time-variation and latent states in any distribution parameters, not only the location. This is the methodology adopted in this chapter.

The remainder of this section extends the score driven dynamic factor model of Creal et al. (2014) in three directions. First, it introduces dynamic factor structures in the scale and shape parameters and discuss their implications in a mixed-frequency setting. Second, it adapts the approach to a nowcasting setting by adding a filtering equation following Buccheri et al. (2021). Third, it relaxes the assumption of independent prediction errors by making use of a copula. To accommodate this feature the dynamic in the model is deliberately disconnected from the observation density: the model becomes *quasi* score driven. Overall the resulting model is able to use correlated related series for updating rapidly point forecasts and, importantly, the uncertainty attached to them as well.

5.2.1 A Mixed-Measurement Approach Relying on a Copula

As in Creal et al. (2014) and Gorgi et al. (2019) each element of the observation vector $y_t = (y_{1,t}, y_{2,t})'$, $t = 1, ..., N$ can have a distinct conditional (or predictive) density such that

$$y_{i,t} \sim f_i(y_{i,t}|Y_{t-1}), \quad i = 1, 2. \tag{5.1}$$

The set $Y_{t-1} = \{z_t, X_{t-1}, \Theta\}$ includes the information available at time $t-1$. Both variables are modelled at a monthly frequency. The GDP series, which is quarterly, shows a quarterly figure in the last month of each quarter with missing values in other months. The vector z_t includes the dynamic states related to location, scale and shape parameters. Θ is a set of fixed parameters such as autoregressive coefficients and factor loadings.

The most straightforward approach to derive the contemporaneous joint log density of the data consists of assuming cross-sectional independence conditional on past observations; this is the strategy followed by Creal et al. (2014) and Gorgi et al. (2019). In this particular case the joint log density is simply the sum of the log marginal densities, such that

$$\log f^*(y_t|Y_{t-1}) = \sum_{i=1}^{2} \delta_{i,t} \log f_i(y_{i,t}|Y_{t-1}), \tag{5.2}$$

$t = 1, ..., N$, where $\delta_{i,t}$ is zero if observation $y_{i,t}$ is missing and one otherwise.

The related series used to improve nowcasts of the target series is chosen on the basis of its high comovement with GDP. While most of this comovement is captured through the common factor in the locations, some residual comovement is likely to remain in the prediction errors, especially round economic downturns like the 2007 financial crisis and the Covid-19 pandemic. Dependence across series in the prediction errors violate the conditional independence assumption on which relies 5.2.

Following the idea presented in chapter 6, the residual dependence in the prediction errors, which have distinct conditional distributions, is captured with a copula. A bivariate copula $C(F_1(y_{1,t}), F_2(y_{2,t}))$ is a joint *cdf* where $F_1(.)$ and $F_2(.)$ are the marginal *cdf*'s. The corresponding joint *pdf* is

$$\begin{aligned} \frac{\partial^2 C(F_1(y_{1,t}), F_2(y_{2,t}))}{\partial y_{1,t} \partial y_{2,t}} &= \frac{\partial^2 C(F_1(y_{1,t}), F_2(y_{2,t}))}{\partial F_1(y_{1,t}) \partial F_2(y_{2,t})} \frac{\partial F_1(y_{1,t})}{\partial y_{1,t}} \frac{\partial F_2(y_{2,t})}{\partial y_{2,t}}, \\ &= c(F_1(y_{1,t}), F_2(y_{2,t})) f_1(y_{2,t}) f_2(y_{2,t}), \end{aligned} \tag{5.3}$$

where $c(F_1(.), F_2(.))$ is the *copula density* and $f_1(.)$ and $f_2(.)$ are the marginal *pdf*'s. Patton (2006) provides a suitable conditional copula theory fitted for modelling the joint distribution of $y_t = (y_{1,t}, y_{2,t})'$ conditional on the past observations $y_{t-1}, ..., y_1$.

Using a copula the log density of the observations at time t becomes

$$\log f(y_t|Y_{t-1}) = \sum_{i=1}^{2} \delta_{i,t} \log f_i(y_{i,t}|Y_{t-1}) + \delta_{1,t}\delta_{2,t} \log c(F_1, (y_{1,t}|Y_{t-1}; \Theta), F_2(y_{2,t}|Y_{t-1}; \Theta)), \tag{5.4}$$

$t = 1, ..., N$, where the dimension of the copula is adapted to accommodate missing values when needed.

This chapter explores the use of two copulae: the Student's t copula and the Gaussian copula. The bivariate Student's t copula is

$$C(u_1, u_2) = t_{2,\nu}(t_\nu^{-1}(u_1) + t_\nu^{-1}(u_2)), \tag{5.5}$$

where $t_{2,\nu}$ is the cumulative distribution function of a bivariate student's t with degrees of freedom set to ν, mean zero and covariance matrix (or standardised dispersion matrix) equal to $R \in [-1, 1]^{2\times 2}$ with ones on the diagonal, and where t_ν^{-1} is the quantile function of a standard student's t with ν degrees of freedom.

When the degree of freedoms of the copula tend to infinity the Student's t copula is reduced to the Gaussian copula:

$$C(u_1, u_2) = \Phi_2(\Phi^{-1}(u_1) + \Phi^{-1}(u_2)), \tag{5.6}$$

where Φ_2 is the cumulative distribution function of a bivariate normal with mean zero and covariance matrix equal to $R \in [-1, 1]^{2\times 2}$ with ones on the diagonal, and where Φ^{-1} is the quantile function of a standard normal distribution.

Finally, the independent copula is retrieved when the dependence matrix in the Gaussian copula is equal to the identity matrix. In this case the copula log density is equal to zero and the total log likelihood is simply equal to the sum of each distribution's log likelihood.

5.2.2 A Score Driven Dynamic for the Unobserved Components

Following Creal et al. (2013) and Harvey (2013) the dynamic in the vector of time-varying parameters z_t comes from the conditional score. However, the score here is not derived from the observation density, but from a constrained version of the latter, thus generating a *quasi* score driven model (see Blasques et al. (2020)). Specifically, the density used for deriving the score is (5.2) which assumes cross-sectional independence (the independent copula). The copula therefore does not affect the dynamic in the model which improves estimation. This point is discussed further in chapter 6, section 6.3.3.

The dynamic equation for the latent states is

$$z_{t+1} = Bz_t + As_t, \tag{5.7}$$

where s_t denotes the scaled first derivative of the log density with respect to the vector z_t:

$$s_t = S_t \Delta_t', \quad \Delta_t = \frac{\partial \log f^*(y_t|Y_{t-1})}{\partial z_t}, \tag{5.8}$$

The unknown parameters in the matrix A are estimated via maximum likelihood along the unknown elements of the matrix B and the initial vector z_1. The matrix A pre-multiplies the scaled score and determines the degree of variation in the unobserved components (their dependence to the dynamic introduced by the score). The matrix S_t is a scaling matrix set to the Moore–Penrose inverse of the expected information matrix

[1] given by

$$\mathcal{I}_t = \mathrm{E}\big[\Delta_t'\Delta_t|Y_{t-1}\big]. \tag{5.9}$$

Since the log density (5.2) relies on conditional independence, the score can be expressed conveniently as

$$\Delta_t = \sum_{i=1}^{2}\delta_{i,t}\Delta_{i,t}, \tag{5.10}$$

where $\delta_{i,t} = 0$ if the observation of series i in period t is missing and one otherwise, and where $\Delta_{i,t}$ is the score of series i. The expected information matrix becomes

$$\mathcal{I}_t = \sum_{i=1}^{2}\delta_{i,t}\mathrm{E}\big[\Delta_{i,t}'\Delta_{i,t}|Y_{t-1}\big] = \sum_{i=1}^{2}\delta_{i,t}\mathcal{I}_{i,t}. \tag{5.11}$$

Using the chain rule yields

$$\Delta_{i,t} = \frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial z_t} = \frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial a_{i,t}}.\frac{\partial a_{i,t}}{\partial z_t} \tag{5.12}$$

where $a_{i,t} = (\mu_{i,t}, \sigma_{i,t}, \alpha_{i,t})'$. The expected information matrix for series i can thus be written as

$$\mathcal{I}_{i,t} = \Big(\frac{\partial a_{i,t}}{\partial z_t}\Big)'\mathrm{E}\Bigg[\Bigg(\frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial a_{i,t}}\Bigg)'\frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial a_{i,t}}|Y_{t-1}\Bigg]\frac{\partial a_{i,t}}{\partial z_t}. \tag{5.13}$$

The score and diagonal element of the expected matrix are shown are shown in appendix E. Following Delle-Monache et al. (2020) and Lucas and Zhang (2016), only the diagonal elements of the expectation are used.

[1] The information matrix needs to be inverted at each step of the recursion because its components are dynamic which considerably slows down estimation. To overcome this technical issue the score driven recursion (hence the log likelihood evaluation as well) is written in C++ (I am grateful to Caterina Schiavoni for her help with this) while the optimisation and the remainder of the analysis is carried in R.

A filtering equation for nowcasting

Unlike state space models, score driven models are fully deterministic conditional on past information. In other words, latent states at time t are defined fully from information in $t-1$ (they are predictable perfectly in $t-1$). This means that there is no room for improvement when new data on the related series for time t are released; one-step-ahead forecasts are not subject to any uncertainty. In a purely forecasting exercise that would not be a problem; Koopman et al. (2016) show that score driven models and non-Gaussian state space models have similar predictive accuracy for a wide range of model specifications. In a nowcasting exercise, however, updating latent states, and thus nowcasts as well, following the release of contemporaneous information is a critical feature; indeed it is the essence of nowcasting models.

There are essentially two strategies for updating nowcasts following the release of contemporaneous information when using a score driven model. The first approach, adopted by Gorgi et al. (2019), consists of leading the related series by one period. While this enables using the related series at t to affect the nowcast of the target series at time t, it is no longer clear what the common factors between the series captures. The other approach put forward by Buccheri et al. (2021), which is adopted in this chapter, consists of using an updating step similar to the updating step of the Kalman filter. When using this strategy score driven models are seen as approximate filters rather than purely predictive models.

The filtering equation (6.4) can be split between an updating step and predictive step as

$$\begin{aligned} z_{t+1} &= B z_{t|t}, \\ z_{t|t} &= z_t + D s_t. \end{aligned} \tag{5.14}$$

Nowcasts are then retrieved by using the filtered vector $z_{t|t}$ rather than one-step-ahead predictions z_t.

The next section presents the components included in z_t (and thus $z_{t|t}$).

5.2.3 Dynamic Factor Models for Location, Scale and Shape Parameters

This section shows how the dynamic factor structure typically employed for location parameters (or conditional means) can be extended to scale and shape parameters. Intuitively, if large prediction errors occur in a series related to GDP, and if these are associated with large prediction errors in GDP, then the dispersion attached to the GDP nowcast should be adjusted accordingly. This is possible directly through the common factor in the scales. On the other hand, modelling dependencies in the shape parameters is useful to capture the asymmetry in prediction errors typically observed at the onset of recessions: While there is an increase in the dispersion of prediction errors, they are likely to be skewed towards negative values. The onset of the Covid-19 pandemic is a good illustration of this fact.

Location parameters are decomposed into a stochastic trend representing idiosyncratic secular changes and a common component to all series which take the following form

$$\lambda^{\mu}_{i,t+1} = \lambda^{\mu}_{i,t} + A^{\mu}_{\lambda i} s^{\mu}_{\lambda i,t}, \quad i = 1, 2, \tag{5.15}$$

$$\pi^{\mu}_{t+1} = A^{\mu}_{\pi} s^{\mu}_{\pi,t}, \tag{5.16}$$

for all periods $t = 1, ..., T$. Antolin-Diaz et al. (2017) demonstrate that random-walk specifications for time-varying parameters are robust to discrete breaks. They also stress the importance of capturing secular changes in economic growth, a finding also discussed by Doz et al. (2020). Common components are usually set as autoregressive processes of order one or two; however, when it comes to modelling extraordinary events like the coronavirus pandemic, suppressing the persistence in the common component of the location helps.

Scale parameters follow a relatively similar model:

$$\lambda^{\sigma}_{i,t+1} = \lambda^{\sigma}_{i,t} + A^{\sigma}_{\lambda i} s^{\sigma}_{\lambda i,t}, \quad i = 1, 2, \tag{5.17}$$

$$\pi^{\sigma}_{t+1} = \phi^{\sigma}_{\pi} \pi^{\sigma}_{t} + A^{\sigma}_{\pi} s^{\sigma}_{\pi,t}, \tag{5.18}$$

for all periods $t = 1, ..., T$. The common factor in the scales captures common volatility shocks, like the one generated by the Covid-19 pandemic, and is specified as a stationary AR(1).

Finally, shape parameters follow:

$$\lambda^{\alpha}_{i,t+1} = \phi^{\alpha}_{\lambda} \lambda^{\alpha}_{i,t} + A^{\alpha}_{\lambda i} s^{\alpha}_{\lambda i,t}, \quad i = 1, 2, \tag{5.19}$$

$$\pi^{\alpha}_{t+1} = \phi^{\alpha}_{\pi} \pi^{\alpha}_{t} + A^{\alpha}_{\pi} s^{\alpha}_{\pi,t}, \tag{5.20}$$

for all periods $t = 1, ..., T$. Unlike in location and scale parameters, the idiosyncratic component in the shape is not a random walk but an AR(1) process. This choice tends to improve estimation. The common factor in the shapes captures simultaneous shifts in the asymmetry of prediction errors, like those typically arising at the beginning of recessions, and is also modelled as a stationary AR(1).

The next section discusses how trends and common components are related to the location, scale and shape parameters.

5.2.4 A Mixed-Frequency Approach for Location, Scale and Shape Parameters

Each series follows the predictive model:

$$y_{i,t} = \mu_{i,t} + \sigma_{i,t}\epsilon_{i,t}, \tag{5.21}$$

for $i = 1, 2,\ t = 1, ..., N$, and where $v_{i,t} = \sigma_{i,t}\epsilon_{i,t}$ is a prediction error (and $\epsilon_{i,t}$ its standardised counterpart) following an arbitrary distribution with time-varying location, scale and shape parameters ($\mu_{i,t}$, $\sigma_{i,t}$ and $\alpha_{i,t}$). These are decomposed into latent states whose dynamics come from the scaled score as outlined in the previous chapter. These latent states include a component common to all variables. However, the data used for estimation are of different nature; while GDP is a quarterly variable, the related series is a monthly variable. It is important to account explicitly for this mismatch in the frequency of aggregation when relating the latent states to the parameters.

Location Parameters

There is a linear relationship between each variable and its location parameter which makes it possible to split the location of a quarterly variable into monthly components and relate them to the location with (A.6). Importantly this step does not require any assumption about the distribution of the unobserved monthly variable.

Monthly locations are modelled as

$$\tilde{\mu}_{i,t} = \lambda^{\mu}_{i,t} + \Lambda^{\mu}_{i}\pi^{\mu}_{t}, \tag{5.22}$$

where the factor loading is constrained to be one for GDP.

Location parameters in quarterly series are related to the monthly model using approximation (A.6) as

$$\mu_{i,t} = \frac{1}{3}\tilde{\mu}_{i,t} + \frac{2}{3}\tilde{\mu}_{i,t-1} + \tilde{\mu}_{i,t-2} + \frac{2}{3}\tilde{\mu}_{i,t-3} + \frac{1}{3}\tilde{\mu}_{i,t-4}, \tag{5.23}$$

while for monthly series it is simply

$$\mu_{i,t} = \tilde{\mu}_{i,t}. \tag{5.24}$$

Scale and Shape Parameters

Diverging from the normal distribution is useful to capture the features of the data in a flexible way, but it also creates challenges in a mixed-frequency framework. Notably, while location parameters can be disaggregated temporally without making any statistical assumptions on the monthly sub-components, that is not possible for scale and shape parameters. This is problematic because the sum of two non-normally distributed variables with known distributions has a distribution which is difficult to evaluate and in many cases unknown. Linking the monthly model to the quarterly model is therefore difficult.

When working in a non-Gaussian framework it is preferable to model the dependencies across series in scale and shape parameters using quarterly models. To do so the related series is aggregated into rolling quarterly observations (quarter-on-quarter deviations observed monthly). This introduces serial correlation in the series which is addressed by modelling monthly location components, as outlined above. The temporal aggregation strategies for modelling jointly quarterly and monthly observations and modelling rolling quarterly observations (quarterly observations observed monthly) are essentially the same; in both cases it is necessary to go back to the underlying monthly model. Chapter 3 shows that when rolling quarterly series are not subject to important measurement errors it is possible to interpolate the monthly path very precisely with (A.6). Hence there should be little loss of information when using rolling quarterly observations instead of monthly observations.

Separately, while locations can take any real value, scale parameters are positive and shape parameters can take values only in (0;1) in the distribution used below. To incorporate these constraints during estimation, trends and common factors are related to scale and shape parameters as

$$\sigma_{i,t} = \exp(\lambda^{\sigma}_{i,t} + \Lambda^{\sigma}_{i}\phi^{\sigma}_{t}), \qquad \alpha_{i,t} = 1/(1 + \exp(\lambda^{\alpha}_{i,t} + \Lambda^{\alpha}_{i}\phi^{\alpha}_{t})), \tag{5.25}$$

with the factor loadings constrained to one for the GDP series ($i = 1$).

The Special Case of a Normal Distribution

In a Gaussian setting it is possible to link monthly variances to quarterly variances using an approximation close to (A.6) as shown below.

If $y_{i,t}$ is conditionally normally distributed, its conditional variance is equal to the variance of the prediction error such that

$$Var(y_{i,t}|Y_{t-1}) = E[(y_{i,t} - E(y_{i,t}|Y_{t-1}))^2] = E(v^2_{i,t}) = Var(v_{i,t}).$$

Assuming that series i is observable quarterly, the quarterly prediction error can be split into its monthly components using approximation (A.6) as

$$v_{i,t} = \frac{1}{3}\tilde{v}_{i,t} + \frac{2}{3}\tilde{v}_{i,t-1} + \tilde{v}_{i,t-2} + \frac{2}{3}\tilde{v}_{i,t-3} + \frac{1}{3}\tilde{v}_{i,t-4},$$

where $\tilde{v}_{i,t}$ is a monthly prediction error. Using this approximation the variance of the prediction error can be decomposed into monthly variances as

$$Var(v_{i,t}) = \frac{1}{9}Var(\tilde{v}_{i,t}) + \frac{4}{9}Var(\tilde{v}_{i,t-1}) + Var(\tilde{v}_{i,t-2}) + \frac{4}{9}Var(\tilde{v}_{i,t-3}) + \frac{1}{9}Var(\tilde{v}_{i,t-4}), \tag{5.26}$$

since, assuming that the monthly errors $\tilde{v}_{i,t}$ are i.i.d, the covariance terms are zero. This step is possible only if the prediction errors are normally distributed. The rolling quarterly scale parameter $\sigma_{i,t}$ can thus be written as a function of the monthly scale parameter $\tilde{\sigma}_{i,t}$ as

$$\sigma_{i,t} = \sqrt{\frac{1}{9}\tilde{\sigma}^2_{i,t} + \frac{4}{9}\tilde{\sigma}^2_{i,t-1} + \tilde{\sigma}^2_{i,t-2} + \frac{4}{9}\tilde{\sigma}^2_{i,t-3} + \frac{1}{9}\tilde{\sigma}^2_{i,t-4}}. \tag{5.27}$$

It is thus possible to specify a monthly model for all scale parameters where the monthly components are related to the quarterly scale parameter of GDP using (5.27) with $\tilde{\sigma}_{1,t} = \exp(\lambda^{\sigma}_{1,t} + \Lambda^{\sigma}_{1}\phi^{\sigma}_{t})$.

5.2.5 The Generalised Asymmetric Student-t Distribution

Each series' conditional density comes from the family of asymmetric student-t (AST) density of Zhu and Galbraith (2010). The distinctive features of the AST are its shape parameter, or skewness parameter, which controls the asymmetry round the central part of the distribution, and its tail parameters which control tail behaviour independently on each side.

The log AST density of observation t of series i takes the form

$$\begin{aligned}\log f_i(y_{i,t}|Y_{t-1}) = &-\ln\sigma_{i,t} - \frac{\nu_{1,i}+1}{2}\ln\Big[1 + \frac{1}{\nu_{1,i}}\Big(\frac{y_{i,t}-\mu_{i,t}}{2\alpha_{i,t}\sigma_{i,t}K(\nu_{1,i})}\Big)^2\Big]\ 1(y_{i,t} \leq \mu_{i,t}) \\ &- \frac{\nu_{2,i}+1}{2}\ln\Big[1 + \frac{1}{\nu_{2,i}}\Big(\frac{y_{i,t}-\mu_{i,t}}{2(1-\alpha_{i,t})\sigma_{i,t}K(\nu_{2,i})}\Big)^2\Big]\ 1(y_{i,t} > \mu_{i,t}),\end{aligned} \tag{5.28}$$

where $\sigma_{i,t}$ is the scale parameter, $\alpha_{i,t}$ is the shape parameter which can take values in

$(0, 1)$, and $\nu_{1,i}$ and $\nu_{2,i}$ are respectively the left and right tail parameters which take positive values. $K(\nu) = \Gamma((\nu + 1)/2)/(\sqrt{\nu\pi}\Gamma(\nu/2))$ ($\Gamma(.)$ is the Gamma function) and $1(x)$ is an indicator variable equal to one if statement x is true and zero otherwise. The distribution is skewed towards positive values if $\alpha_{i,t} < 0.5$ and towards negative values if $\alpha_{i,t} > 0.5$. When the tail parameters are constrained to be very large and the skewness parameter to 0.5 the AST is equivalent to a (scaled) normal distribution.

Figure 5.2 illustrates the effects that scale, shape and tail parameters have on the density function. The solid red line shows the AST density when $\sigma = 1$ while other parameters are constrained to be Gaussian ($\alpha = 0.5$, $\nu_1 = \nu_2 = \infty$). The dotted blue line shows the AST density when either the scale, tail or shape parameters vary. The density at the location changes when the scale parameter is altered but is independent to variations in the tail and shape parameters. This is a particular feature of this version of the AST distribution (given in equation (5) of Zhu and Galbraith (2010)) where the random variable is scaled with $B_{i,t} = \alpha_{i,t}K(\nu_{i,1}) + (1 - \alpha_{i,t})K(\nu_{i,2})$.

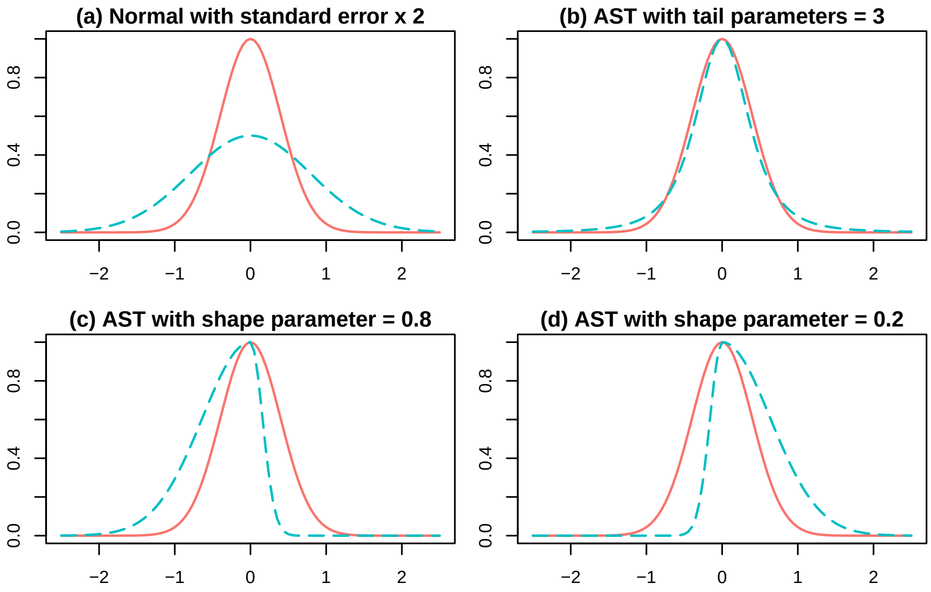


Figure 5.2: Illustration of the density functions from AST distributions with different set of parameters. With shape parameter = 0.5 and tail parameters = ∞ the AST distribution reduces to a (scaled) normal distribution (red solid line).

Graph (a) shows the effect of increasing the scale parameter, specifically to $\sigma = 2$.

The dispersion increases symmetrically on both sides and can be interpreted as a general (or symmetric) increase in forecasting uncertainty when the model is applied to conditional GDP growth. Graph (b) shows the effect of lowering both tail parameters to three. The probability in the tails increases which can be used to account for extreme events. Graph (c) shows the effect of negative skewness ($\alpha = 0.8$) on the density function. The central part of the density becomes skewed heavily toward negative values. This statistical behaviour should be especially useful to capture the onset of recessions when coupled with an increase in dispersion: while general economic uncertainty increases, the likelihood of a positive outcome decreases. The real-time analysis of US GDP in the first quarter of 2020 shown in section 5.7 validates this intuition. Finally, graph (d) shows the opposite behaviour, that is positive skewness ($\alpha = 0.8$), which should be useful in the early part of economic recovery following the Covid-19 recession.

Figure 5.3 illustrates the response function of the scaled score for location, scale and shape parameters with the prediction error as input.[2] These are useful for illustrating two important properties of the scale score. First, introducing a skewness parameter yields a discontinuity round zero. For instance, a distribution skewed towards positive values downweights the effect of positive prediction errors and amplifies the effect of negative ones. Secondly and most importantly, the scale score downweights the effect of large prediction errors when tail parameters are small (low rates of decays in the tails). The Gaussian and skewed models, on the other hand, yield linear or increasing responses when prediction errors increase.

[2] The components of the score and information matrix are shown in appendix E.

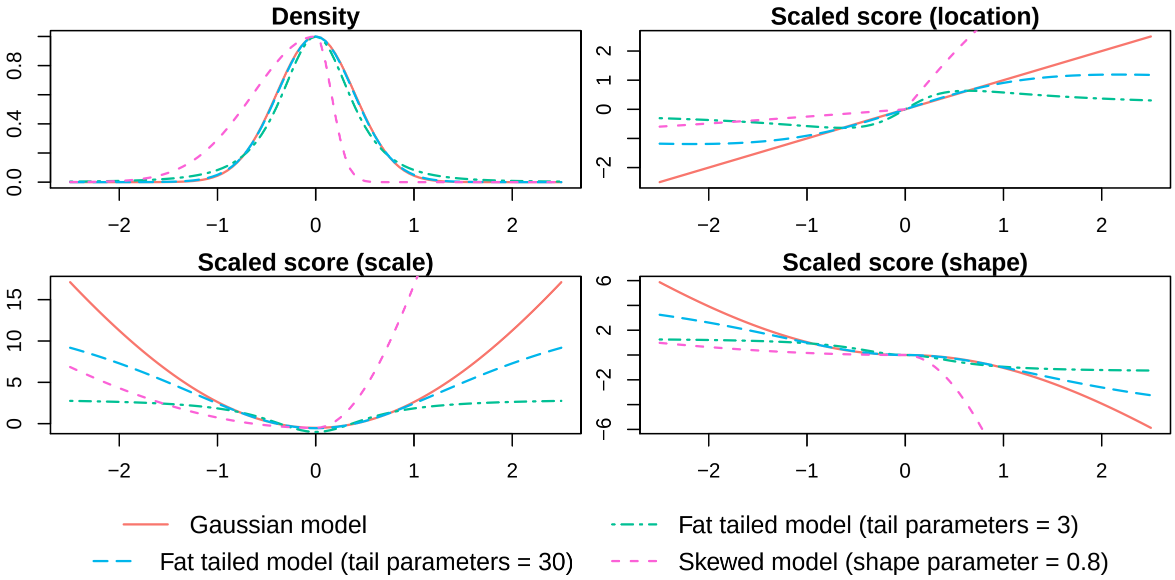


Figure 5.3: Plots of the density and score functions for an asymmetric student-t (AST) distribution with $\sigma = 1$. X-axes show prediction errors. With shape parameter $\alpha = 0.5$ and tail parameters $\nu_1 = \nu_2 = \infty$ the AST is equivalent to a (scaled) normal distribution.

Low tail parameters can complicate the identification of turning points

Downweighting the effect of large prediction errors is usually a desirable feature of score driven models because it leads to a more robust estimation. However, economic crises, and the Covid-19 period in particular, are examples of cases where outliers most likely have important and long-lasting effects on means and variances of time series. If the related series' distributions are allowed to have fat tails, the effect of large swings in the economic activity on the time-varying parameters are likely to be downweighted. Consequently the model might not capture turning points in the economic activity and the concurrent increase in the dispersion of prediction errors in a timely way.

To investigate thoroughly the effects of fat tails on estimation and forecasting performance in recessionary episodes, the empirical analysis compares a model with unconstrained tail parameters with models forced to have Gaussian tail parameters in the related series.

5.3 A Weighted Maximum Likelihood Estimator

The estimation strategy relies on the weighted maximum likelihood method of Blasques et al. (2016), notably applied in a score driven framework by Gorgi et al. (2019). It accounts explicitly for the fact that, while related series are used for estimation, the primary objective is to forecast the target series (GDP growth), not the related series; these are used solely to improve GDP nowcasts.

In a misspecified setting, the parameters that maximise the total log likelihood (5.4) are not necessarily those that maximise the log likelihood of GDP. This issue is even more prominent in a mixed-frequency framework where GDP is observed once every three months whereas the related series is observed monthly. Indeed, when an observation is missing, the series it relates to has no impact on the model's log likelihood. The log likelihood associated to GDP has consequently a lower weight on the total log likelihood than those of the related series.

The vector of unknown parameters Θ which includes the marginal distributions' parameters, the copula parameters, the initial values of the time-varying parameters, the autoregressive coefficients and the gains is estimated as

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{t=1}^{N} \log f^{w}(y_t|Y_{t-1}), \tag{5.29}$$

where $\log f^{w}(y_t|Y_{t-1})$ is the weighted log likelihood defined as

$$\begin{aligned}\log f^{w}(y_t|Y_{t-1}) = \delta_{1,t}\log f_1(y_{1,t}|Y_{t-1}) + \delta_{1,t}\delta_{2,t}\log c(F_1,(y_{1,t}|Y_{t-1};\Theta), F_2(y_{2,t}|Y_{t-1};\Theta))\\ + W\delta_{2,t}\log f_2(y_{i2,t}|Y_{t-1}).\end{aligned} \tag{5.30}$$

The weight W is applied to the related series' marginal log likelihood and decreases its contribution to the total log likelihood. It cannot be estimated alongside the other parameters and Blasques et al. (2016) suggest selecting it via cross-validation techniques. Alternatively, Gorgi et al. (2019) set the weight to zero. This complicates the identification of the scale parameters in the related series; a problem they overcome by modelling a unique scale parameter for all series. This would be problematic here since the modelling approach proposed relies on a flexible specification for scale parameters. Furthermore, in their out-of-sample nowcasting exercise Gorgi et al. (2019) do not find that the score driven approach yields a clear benefit when the contribution of the indicator series is null.

In this chapter the weighted maximum likelihood approach is used to offset the implicit downweighting of the GDP series's contribution to the total log likelihood coming from its lower frequency of observation compared to the related series. For each observation of GDP, three observations of the related series are available. This requires setting the weighing factor to three. The log likelihood contribution of GDP is not increased further because that might deteriorate excessively the model's capacity to predict related series in real time. This could yield to prediction errors losing their economic meaning and with it their ability to indicate periods of economic depressions and uncertainty, which is the focus of this chapter.

5.4 Real-Time Data on US Economic Growth and Employment

This study is centred on quarterly GDP which is the leading measure of economic growth and the data are taken from the United States. The estimate of GDP analysed is the *Advance Estimate*, the most rapid estimate of GDP in the US.

One monthly series related to economic growth is used to improve GDP nowcasts:

the index of total aggregate weekly hours.

Both series are occasionally subject to benchmark changes which affect the entire series, such as changes in indexation. To deal with these changes they are taken in first differences in logs. The sample includes data from January 1973 up to June 2020. The series are illustrated in figure 5.4 using the vintage available at the end of July 2021.

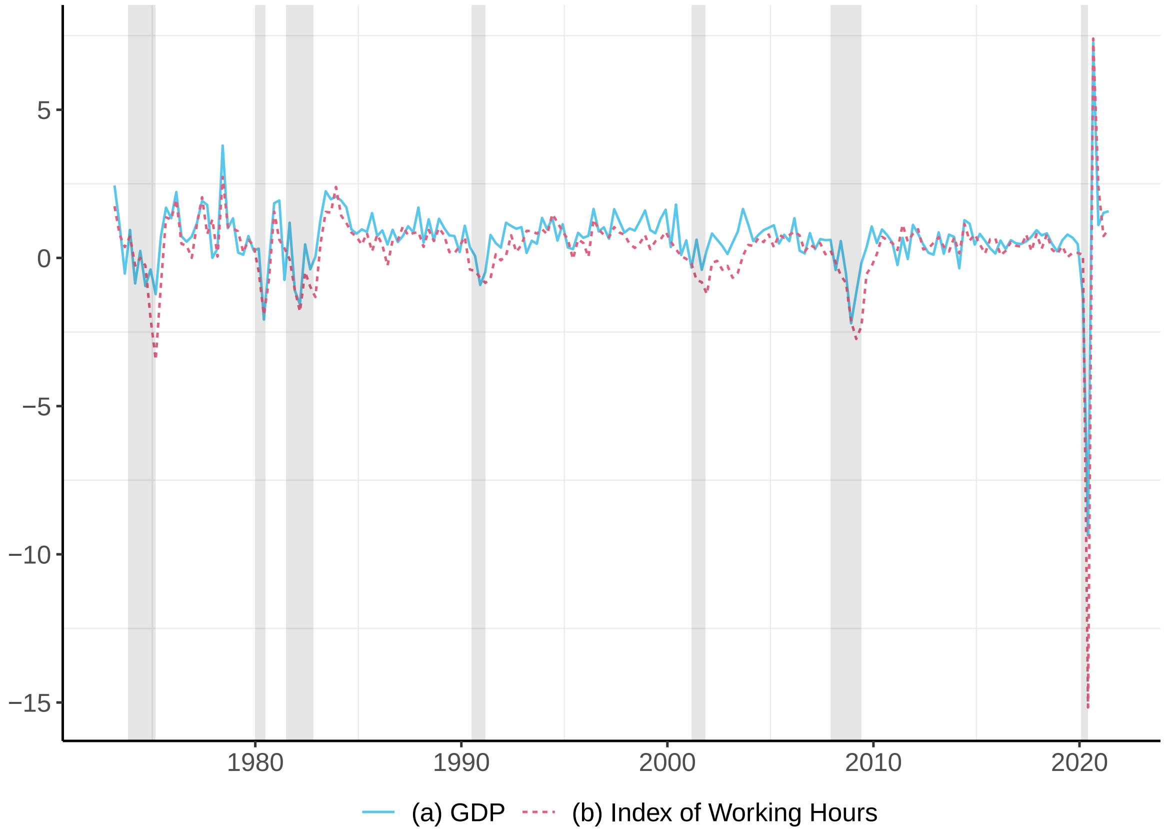


Figure 5.4: Quarterly data (calendar quarters) from January 1973 up to June 2021. The index of total aggregate weekly hours is aggregated into quarterly figures for comparison with GDP. Data from the most recent vintage. Shaded areas indicate the US recessions classified by the NBER.

There is a clear comovement in the data which is confirmed by the correlation coefficients shown in table 5.1. Since recessions are of particular interest to forecasters, cross-correlation coefficients are shown for recessionary episodes as classified by the NBER and normal times. The series are highly correlated during recessions which suggests that employment data should be helpful in capturing turning points in economic activity. In normal times both series remain closely correlated, albeit on a lower magnitude.

Table 5.1: Correlation across GDP and the index of total aggregate weekly hours.

Levels		Absolute deviations	
Normal Times	Recessions	Normal Times	Recessions
0.780	0.930	0.760	0.940

Note: The data are in taken in first differences in logs.

The second half of table 5.1 shows correlation coefficients of absolute deviations, which are a measure of short-term volatility. It is better to analyse absolute deviations and not squared deviations because the latter would give too much weight to outlying observations. The picture is similar to levels; short-term volatilities are highly correlated during recessions while remaining significantly correlated in normal times. This suggests that modelling dependencies in scale parameters should be helpful.

To filter short-term fluctuations and get a better idea of the underlying trends in the mean and volatility of each series, Figure 5.5 shows six-year moving averages of the levels and squares of the series. While there is a strong similarity in the underlying trends of both levels and squares, there are some periods of deviations. These deviations can be captured with the idiosyncratic trend in each parameter.

GDP and the index of total aggregate weekly hours are systematically revised over time because the data used for their compilation accrue gradually. Therefore, to investigate the forecasting performance of the model in real time it is necessary to estimate the model recursively using successive vintages of the data. Such vintages are produced by the Federal Reserve Bank of Philadelphia.

The index of total aggregate weekly hours is the most timely series. It is released at the beginning of each month and relate to the previous month. GDP is released at the end of the month following the quarter it relates to. This yields four rounds of revisions in between GDP releases, that is four nowcasting steps. Table 5.2 illustrates the vintages of the data available at different steps from early April to early May 2020.

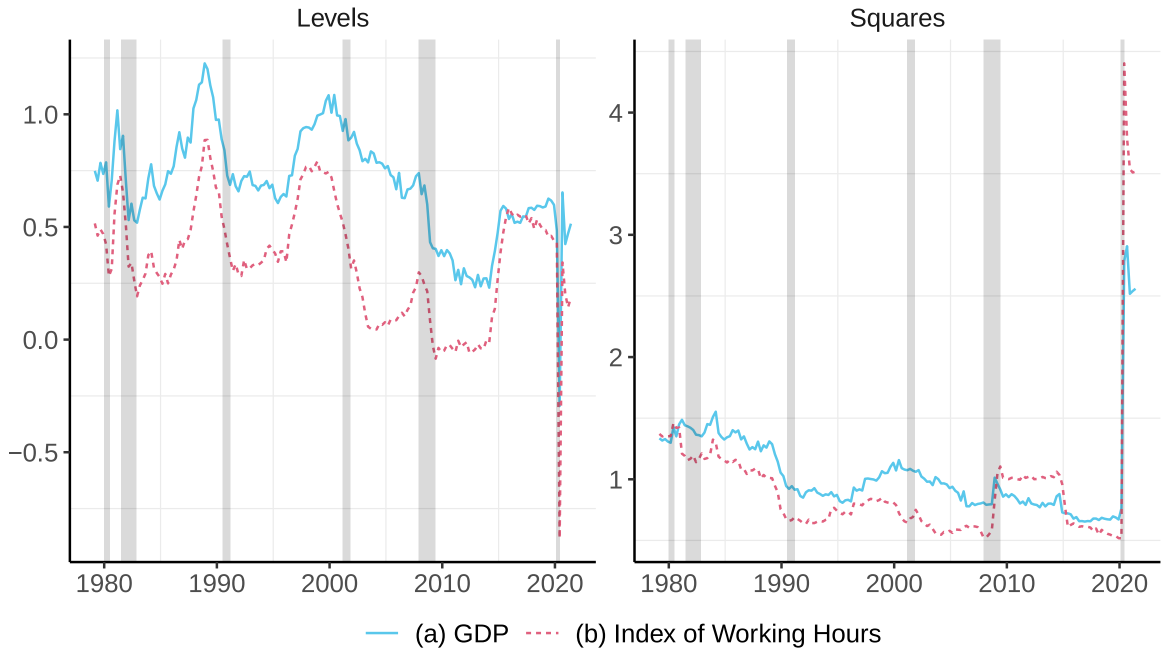


Figure 5.5: Six-year moving average of levels and squared quarterly figures (calendar quarters only). Data from January 1973 up to June 2021. Shaded areas indicate the US recessions classified by the NBER.

The first nowcasting step is at the end of the first month in the quarter, when the previous-quarter GDP figure is released. The second step is at the beginning of second month in the quarter when monthly figures for the index of total aggregate weekly hours are released. The next two steps are approximately one month apart, when figures for the index of total aggregate weekly hours are released.

5.5 Six bivariate models

Six bivariate models are compared to investigate the potential gains of modelling common factors in scale and shape parameters on nowcasting performance. All models feature a dynamic factor model in their locations, but specifications for the conditional distributions and scale and shape parameters vary.

The most flexible specification is model (a) which is a rolling quarterly model with stochastic volatility and shape. Through the common factors in these parameters the

Table 5.2: Illustration of the vintages used for the real-time estimation exercise, from early February to late May 2020.

	Vintages							
	Early February		Early March		Early April		Late April	
Series	GDP	IWH	GDP	IWH	GDP	IWH	GDP	IWH
Jan 20		×		×		×		×
Feb 20				×		×		×
Mar 20						×	×	×

Note: GDP = Gross Domestic Product (Advanced estimate), IWH = Index of Total Aggregate Weekly Hours.
× indicate a figure relating to the month specified in the first column; for GDP it is a quarterly figure relating the past three months.

related series can be used in a timely way to adapt the dispersion and shape of US GDP growth density nowcasts. Each series follows a distinct Skew t-distribution. As discussed in section 5.2.4, modelling a common factor in the shape parameters requires aggregating the related series into rolling quarterly figures. Model (b) is a constrained version of model (a) where series are modelled as Skew normal instead of Skew t.

Model (c) is a stochastic volatility model with no skewness. The employment series is used monthly and follows a Student's t-distribution. GDP growth, on the other hand, is constrained to follow a normal distribution so that its scale parameter can be disaggregated temporally using (5.27). Model (d) is constrained version of model (c) where the employment series is constrained to be Gaussian. The relative performance of models (a) and (b) compared to models (c) and (d) thus provides an indication of the benefit stemming from modelling a common factor in shape parameters. Finally, model (e) and (f) are constant volatility versions of model (c) and (d) respectively.

In the three models discussed so far the tail parameters are constrained to be Gaussian. As discussed in section 5.2.5, this puts more attention on sharp movements in the related series. However, the Covid-19 pandemic and the extreme effect it has had on our economies has revived the debate around fat tails in forecasting. Therefore, to understand the effect of tat tails on nowcasting performance, model (d) extends model

(c) to Student's t related series. Table 5.3 summarises the different model specifications.

5.6 In-Sample Analysis and Copula Benefit

In this section all six models are estimated using the latest available vintage in this study (July 2021). While this analysis in-sample cannot be used to infer the models' relative performances in real time, it serves two purposes. First, it is useful for illustrating the benefit of using a copula which capture cross-sectional dependencies in the prediction errors. Second, it is appropriate for studying the profiles that dynamic parameters take over time, especially scales and shapes.

Table 5.4 shows estimation results and information criteria resulting from estimating each model. The total log likelihood of the model when using a copula is also compared to the total log likelihood when assuming conditional cross-sectional independence (independence copula). When comparing the total log likelihoods and the information criteria across models it is important to bear in mind that the rolling quarterly model (a) and (b) cannot be compared to other models because they have different data: the related series is rolling quarterly whereas it is monthly in other models. The log likelihood attached to GDP, however, is comparable across all models.

Fat tails and stochastic volatility matter greatly

Introducing separately stochastic volatility and fat tails generate very large gains in the model log likelihood which can be observed by comparing the benchmark model with the Student's t and stochastic volatility models. As expected, the model featuring both features is favoured by all information criteria. However, while the Skew t rolling quarterly model with stochastic volatility and shape yields a significant improvement in the total log likelihood, the log likelihood attached to GDP is lower than with the Skew normal model.

Table 5.3: Description of the models

Model	Label	Conditional distributions	Specifications for location, scale and shape parameters	Data
(a)	Rolling quarterly Skew t model with stochastic volatility and shape	Skewed t for each series	Dynamic factor models for locations, scales and shapes.	Quarterly GDP; Rolling quarterly IWH
(b)	Rolling quarterly Skew normal model with stochastic volatility and shape	Skewed normal for each series	Dynamic factor models for locations, scales and shapes.	Quarterly GDP; Rolling quarterly IWH
(c)	Monthly t model with stochastic volatility	Gaussian for GDP growth Student's for IWH	Dynamic factor models for locations and scales. Constant shape parameters.	Quarterly GDP; monthly IWH
(d)	Monthly normal model with stochastic volatility	Gaussian for GDP growth and IWH	Dynamic factor models for locations and scales. Constant shape parameters.	Quarterly GDP; monthly IWH
(e)	Monthly t model	Student's t for GDP and IWH	Dynamic factor models for locations. Constant scale and shape parameters.	Quarterly GDP; monthly IWH
(f)	Monthly normal model	Gaussian for GDP and IWH	Dynamic factor models for locations. Constant scale and shape parameters.	Quarterly GDP; monthly IWH

Note: The skewed normal distribution here is derived by constraining the AST distribution to have Gaussian tail parameters ($\nu_1 = \nu_2 = \infty$). The skewed Student-t is derived by constraining the AST distribution to have a unique tail parameter ($\nu_1 = \nu_2$). The Student-t is derived by constraining the AST distribution to have Gaussian shape and tail parameters ($\alpha = 0.5$ and $\nu_1 = \nu_2 = \infty$). IWH = Index of Total Aggregate Weekly Hours.

Table 5.4: Comparison of the model specifications using the full-sample results.

Model	Log Lik.	Log Lik. indep.	Log Lik. GDP	$\hat{\theta}$	$\hat{\nu}_{cop}$	AIC	BIC
(a) SVS-t	-186.8	-203.09	-163.19	0.43	13.13	427.6	553.27
(b) SVS	-206.14	-230.7	-155.68	0.44	8.56	462.29	578.64
(c) SV-t	-261.15	-271.38	-149.09	0.17	5.53	556.3	635.42
(d) SV	-265.59	-281.04	-159.22	0.23	8.99	563.18	637.64
(e) t	-303.7	-322.06	-181.27	0.17	2.49	631.39	687.24
(f) Benchmark	-406.71	-436.48	-226.92	0.18	0.58	833.43	879.97

Note: Log Lik. refers to the model's total log likelihood when using a copula for estimation. Log Lik. indep. refers to the model's log likelihood when using the independence copula for estimation. Log Lik. GDP refers to the log likelihood of the GDP series only. The total log likelihoods and information criteria of model (a) and (b) are not comparable with the other models (different data). AIC $= -2 \times$ Log Lik. $+ 2 \times p$; BIC $= -2 \times$ Log Lik. $+ \log N \times p$; N is the number of observations and p the number of parameters estimated.

Using a copula yields to a large increase in the log likelihood

Using a copula yields a very large increase in the total log likelihood for each of the model's specification, suggesting important cross-sectional dependencies in the prediction errors. The copula comes at the expense of introducing only two additional parameters: the copula dependence parameter θ and the number of degrees of freedom attached to the Student's t copula ν_{cop}. The dependence across prediction errors is greater in the rolling quarterly model with a dependence parameter of about 0.4 compared to round 0.2 for the models featuring a monthly related series.

The common scale component provides a consistent historical picture of forecasting uncertainty

The left panels of Figure 5.6 show the estimated stochastic scale of GDP growth alongside the scale common factor. The estimates are retrieved from model (a) which features Skew t series. The scale parameter shows an overall decrease in volatility starting in the second half of the eighties which has been documented extensively in the macroeconomics literature since McConnell and Perez-Quiros (2000). Separately, large spikes occur during the recessions triggered by the 2007 financial crisis and lately by the coronavirus pandemic. These are largely coming from the common component. While the common component at the time of the financial crisis reaches levels not seen since the seventies and early eighties, its level during the Covid-19 induced recession is unprecedented. Overall the common volatility factor carries economic meaning and gives a consistent picture of forecasting uncertainty historically.

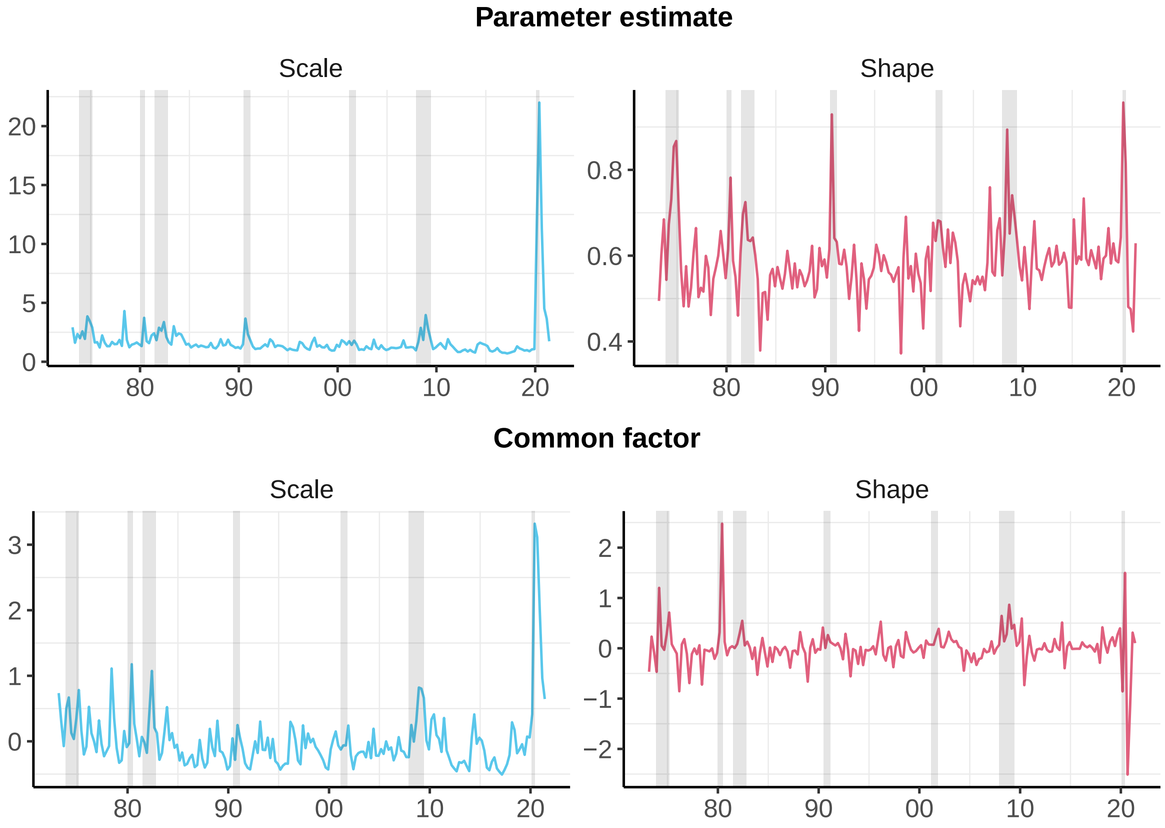


Figure 5.6: Scale and shape parameters of GDP growth and scale and shape common factors (quarterly figures). Estimates from estimation with full sample. Shaded areas indicate the US recessions classified by the NBER.

Downturns are associated with rising negative skewness and recovery with positive skewness

The right plots of figure 5.6 show the shape parameter of GDP growth and shape common component, again resulting from model (a). The increase in forecasting uncertainty when entering recessions discussed above is generally associated with a sharp increase in the skewness parameter: the probability of negative prediction errors increases while positive prediction errors become less likely. This movement in the skewness parameter is then reversed during recoveries. The picture of the shape parameter over time is consistent with the findings of Carriero et al. (2020) and Delle-Monache et al. (2020) who show that, while statistical evidence for skewness in GDP growth is generally weak, this masks an erratic behaviour with periods of positive and negative skewness.

5.7 Real-Time Exercise Covering the Pandemic

The six models are estimated recursively using real-time data since the beginning of the pandemic. The first vintage used for estimation is the March 2020 vintage where employment data for February 2020 have just been published. This is effectively the last month before the effect of the pandemic started to emerge in the data. The model is re-estimated every time new observations are added to the model. It is thus possible to investigate whether capturing dependencies across series in shape parameters is useful for producing accurate density nowcasts of US GDP growth.

Deriving multi-step ahead density nowcasts

The GDP density nowcast in the last step of the nowcasting window is directly given by the one-step ahead prediction error density of GDP at the predicted location. But for earlier steps it is important to account for the uncertainty induced by missing values in the related series. This is done by drawing vectors of observations for each period with missing observations and using the score driven recursion to retrieve time-varying parameters in the next period, which are then used to generate new observations. Specifically, if the target is the GDP nowcast in period t, but the related series is missing from $t-1$, then the one-step ahead joint density of the related series at $t-1$ are used to draw prediction errors centred round the location estimate in $t-1$. The score driven recursion is then applied separately on each prediction to retrieve sets of scale, shape and location parameters for period t. These new parameters yield one-step ahead densities which are used to draw prediction errors round the locations in t. Eventually, each vector of prediction errors in t yields a GDP nowcast for t through the filtering step of the score driven recursion. The density nowcast is given by the empirical density attached to these GDP nowcasts.

Figure 5.7 shows the real-time nowcasted conditional mean and its associated 90% confidence interval for each model. There are four nowcasting steps per quarter. The gray line shows the first release of GDP. First, the models with stochastic volatility and both stochastic volatility and shape do significantly better than the model with constant scale and shape. This is most striking during the recession and the first quarter of the recovery. The model with constant scale and shape and Student's t series do especially poorly; the conditional mean and associated uncertainty barely adjust during the pandemic. In a score driven setting, fat tails can hinder the ability of a model to adjust quickly to the data because the score downweights the effect of outlying observations, as is illustrated in figure 5.3. However, when the scale and shape parameters are allowed to vary this is less of an issue.

While the models with stochastic volatility (c) and (d) yield better forecasts during the second quarter of last year, the models with stochastic shapes are the only models which capture the onset of the pandemic. This is more clearly visible in figure 5.8 which shows the density nowcasts derived at the last step of the nowcasting window, that is approximately four weeks before the release of the first estimate GDP growth. The models with stochastic shape show a clear negative skewness, indicating an increase in the probability of negative prediction errors.

Finally, the density nowcasts are compared across models using the average log score, which is given by the nowcasted log density round the observed value (the first

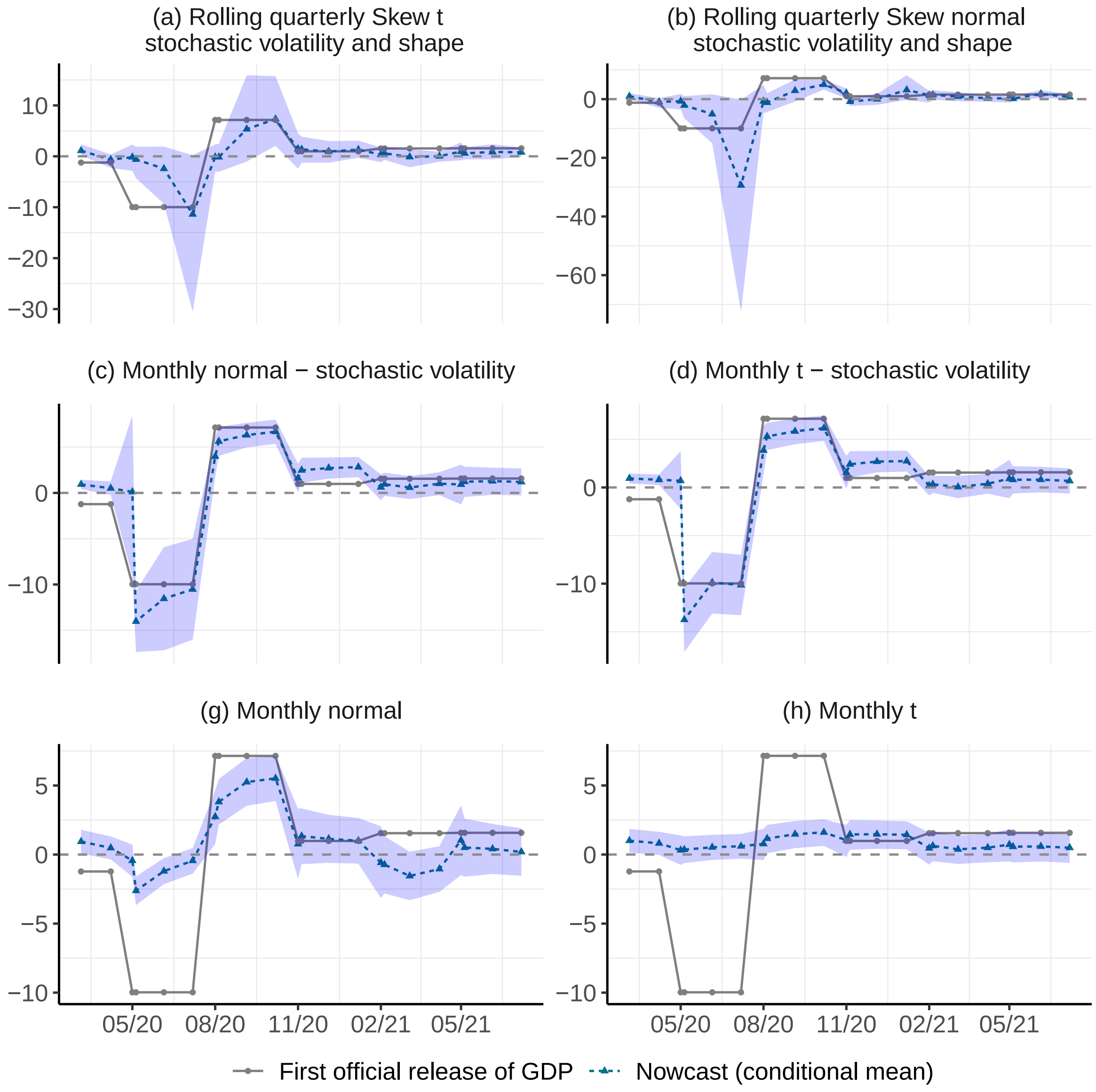


Figure 5.7: Real-time US GDP nowcast at each step of the nowcasting window. Quarterly estimates corresponding to calendar quarters.

estimate of GDP). The better the probabilistic forecast, the greater the log score. The point forecasts, given by the predicted conditional means, are evaluated using the mean absolute error. Figure 5.9 shows the average log score and mean absolute error at each nowcasting step during the quarter. Step four follows the release of the previous-quarter GDP figure, while step one (the last nowcasting step) follows the release of employment data for the last month of the quarter. The results are split between the recession phase (the first two quarters of 2020) and the recovery phase (the following four quarters).

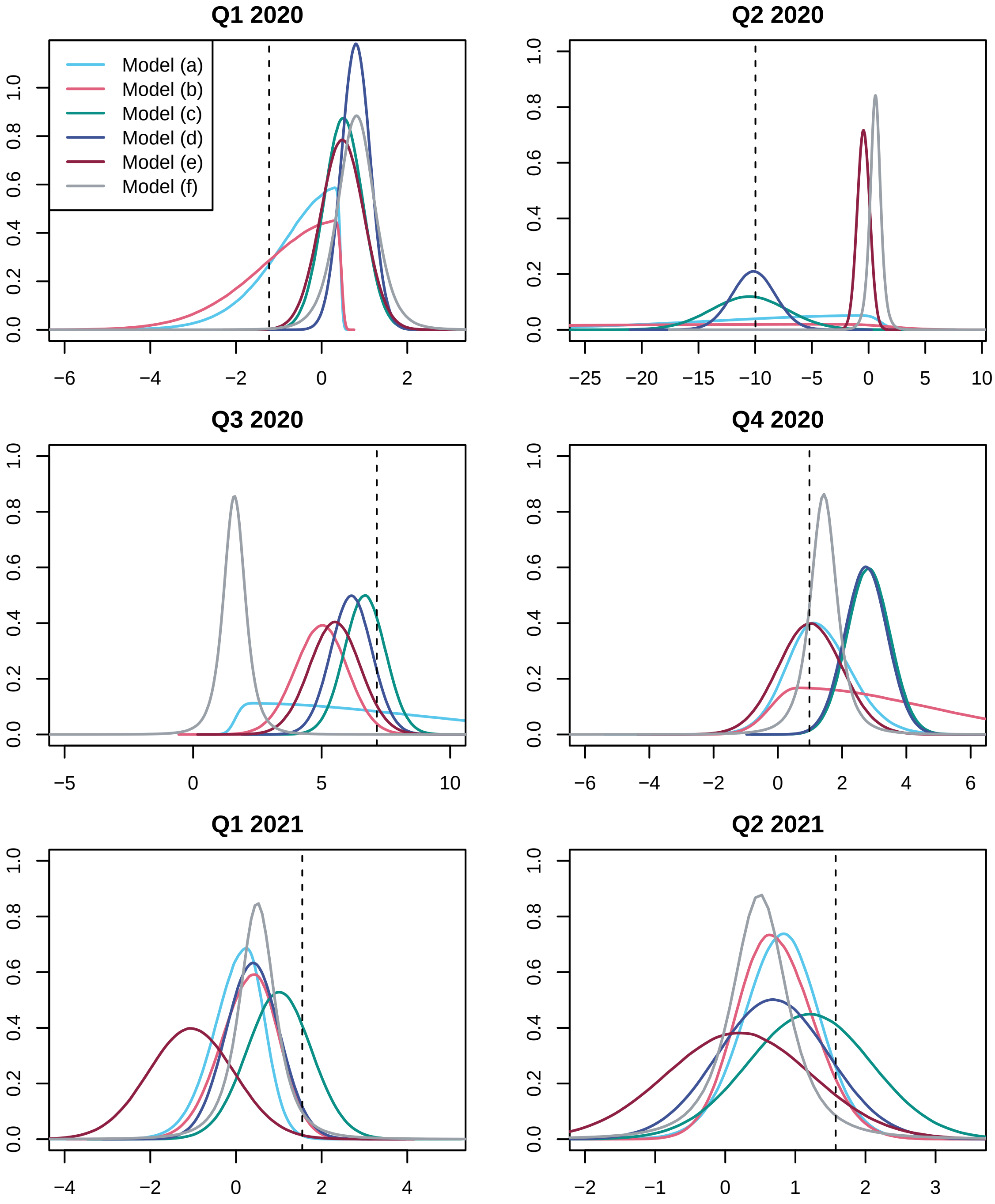


Figure 5.8: Real-time density nowcasts of US GDP at the last step of the nowcasting window (about a month before the first official release of GDP growth). The dashed line indicates the published figure.

As more data accrue during the quarter the average log score should increase monotonically while the mean absolute error should decrease. However, while predictions are generally better towards the end of the nowcasting window, they do not exhibit a monotonic improvement. This can be explained by the relatively low numbers of

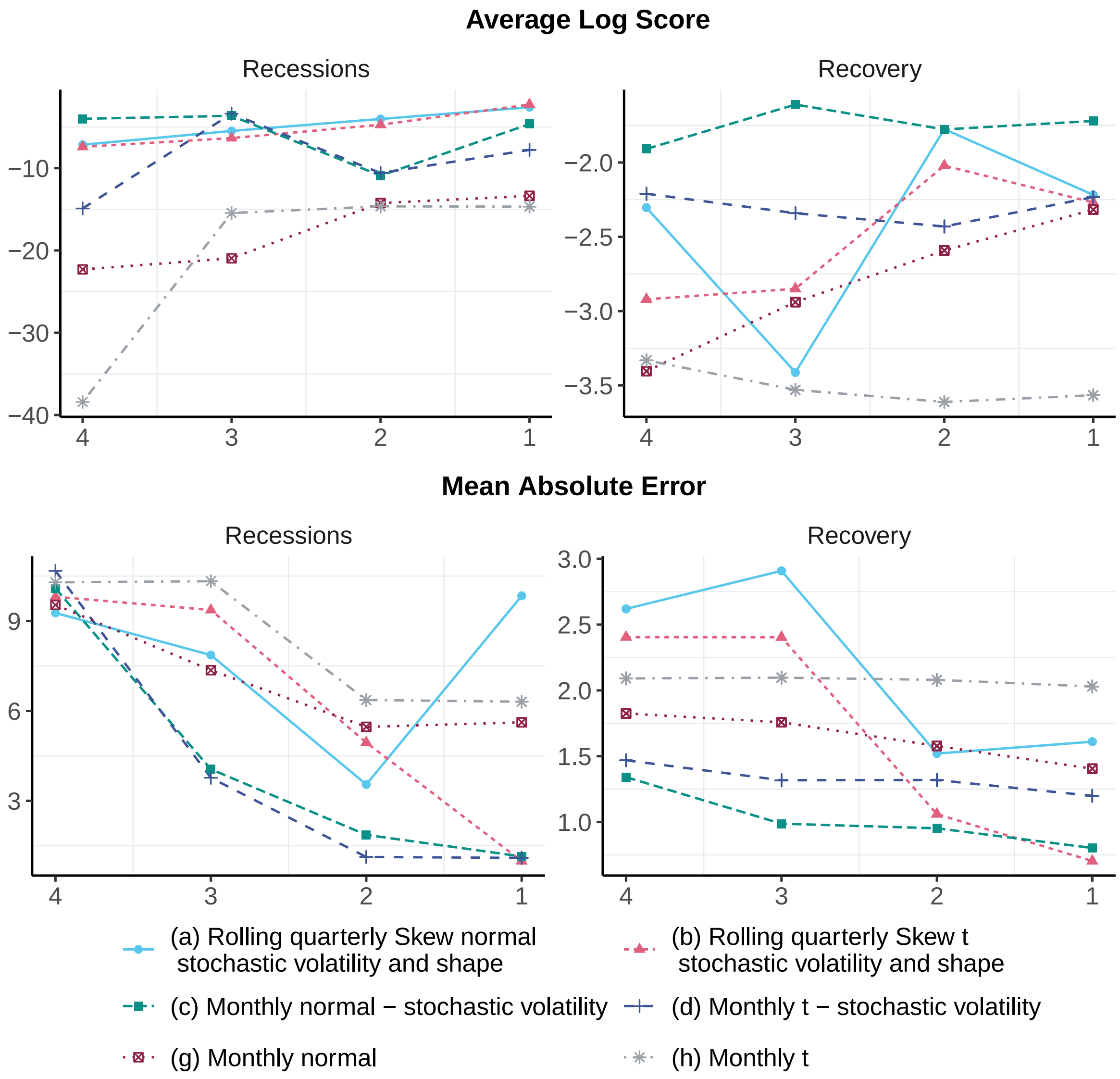


Figure 5.9: Average log score. Higher values imply better density forecasts. x-axes show the numbers of steps towards the release of the GDP number in the nowcasting window. New data are released at each step.

series modelled and the resulting low numbers of release dates during the quarters. Nevertheless, the average log score and the mean absolute error remain useful to compare the models' predictive capacities.

In line with the previous results, the models with stochastic shape tend to yield better probabilistic predictions during the recession. This is especially apparent towards the end of the nowcasting window. This result is driven by the singular ability of the models with stochastic shapes to capture the beginning of the pandemic. The models featuring only stochastic volatility, however, do better during the recovery phase. Here using monthly employment data as opposed to rolling quarterly figures seems to matter most. It is therefore not possible to find a model which ranks best during each period and nowcasting step, suggesting that combining models would be beneficial.

5.8 Conclusion

This chapter shows how stochastic volatility and shape may be modelled across heterogeneous series sampled on different frequencies and released asynchronously. Using score driven techniques it provides an novel alternative approach for nowcasting besides dynamic factor models and MIDAS regressions where stochastic shape parameters have not yet been explored. The empirical application in pseudo real time using US economic growth data shows that modelling scale and shape common factors yields better density nowcasts. The increase in volatility observed during recessions is associated with increased negative skewness. Capturing this movement in the shape of the conditional distribution of GDP growth is critical for predicting accurately the onset of the coronavirus recession.

Chapter 6

A timely indicator of migration net flow using google search queries on national newspapers

6.1 Introduction

So far this thesis has tackled challenges related to the measurement and nowcasting of economic growth using filtering methods, specifically state space and score driven models. This chapter makes use of similar techniques but explores a different problem: measuring migration net flows in a timely way.

Measuring migration flows is a complex and expensive task. Each country exploits a range of different data sources, composed of surveys and administrative data, to measure migration into its territory. Gathering and exploiting these data is typically time-consuming. Consequently migration statistics are not very timely or frequent. Publishing migration statistics several months after the end of the quarter in question is considered rapid.

To produce a timely and high-frequency indicator of migration net flows this chapter explores the use of Google newspaper queries. We regularly follow the news related to the places we are attached to, mainly the country we come from, and the main way to do so is to read national newspapers from that country. Hence, by observing the relative popularity of, for example, newspapers from New Zealand in Australia, it should be possible to infer the net flow of New Zealanders in Australia. Google Trends offers a way to do just that.

Google Trends allows users to visualise the popularity of search queries over time and across countries. To extract a timely indicator of cross-border movements from these data there are conceptual and statistical challenges to solve. Conceptually, variations in the popularity of a newspaper query can be due to many factors i.e. a fall in searches related to *The New Zealand Herald* (a newspaper from New Zealand) in Australia does not necessarily mean that New Zealanders have been leaving Australia. First, it might be that *The New Zealand Herald* has become less popular compared to other newspapers from New Zealand. Second, the popularity of online newspapers compared to new sources of information, such as social media, varies over time. Third, Google normalises its series to account for the total number of searches in a given time period and region.

Accounting for changes in the popularity in-between national newspapers is straightforward: Google Trends lets users visualise the popularity of more than one keyword in a single series i.e. it is possible to visualise the combined popularity of both *The New Zealand Herald* and *Stuff* (another popular news website from New Zealand). Accounting for the other two exogenous sources of variations - the popularity of online newspapers compared to other sources of information and the relative popularity compared to all searches - is more complicated. For these I propose to compare queries searched in a given country with the same queries searched globally. Hence, if the popularity of newspapers from New Zealand increases in Australia but remains constant globally, one can infer that the number of New Zealanders in Australia has increased.

Concerning the statistical problem, google queries are subject to important seasonal movements and noise; two features illustrated in Figure 6.1. This plot shows Google Trends data corresponding to a set of newspapers from New Zealand. The red line shows searches in Australia while the blue line shows searches from everywhere. It is difficult to get a sense of the underlying trends behind the figures because of the seasonality and erratic movements. Seasonality could arise from people reading less news during summer holidays or at Christmas, for instance. Regarding noise, events like national elections or natural disasters can introduce large one-off movements in the data. Seasonality in newspapers reading and noise do not reflect underlying changes in the number of people searching online newspapers and therefore should be discounted.

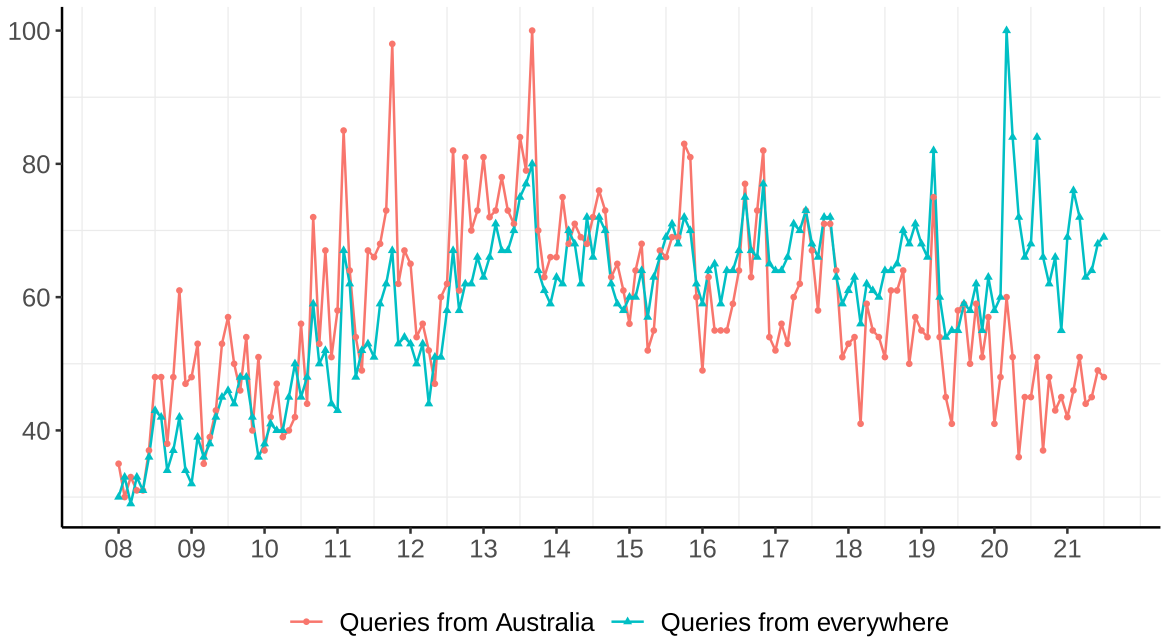


Figure 6.1: Google Trends data of national newspapers from New Zealand. Data extracted with keywords ”nz herald + stuff nz” on the 28th August 2021.

Separating trends from seasonality and noise can be achieved through filtering methods where the data are modelled as a combination of unobserved components. The identification of the trend reflecting cross-border movement is then achieved by modelling jointly the country-specific series and the global series. The trend in the global series must also affect the country-specific series because it is driven by the general popularity of the chosen set of newspapers. Thus, the divergence in the two series comes from a second component, or trend, which affects only the country-specific data and reflects cross-borders movements.

Gaussian state space models are the most popular approach for filtering in economics. However, Google Trends data tend to be very erratic which complicates the implementation of state space models relying on normally distributed data. As an alternative, this chapter exploits score driven techniques (Harvey (2013) and Creal et al. (2013)) as in chapters 5 and 4.

Separately, although the country-specific and world series exhibit a degree of comovement, their conditional distributions can diverge significantly in terms of kurtosis and skewness. This is a general feature of Google Trends series which can diverge greatly in their time-series properties; some might be characterised best by a fat tailed distribution while others are represented better with a skewed distribution. This is problematic because most models rely on the data following a multivariate distribution where tail and shape parameters are common to all series.

To model openly the heterogeneous nature of the data, this chapter presents a new approach based on copulae specifically suited for modelling jointly series which are related but characterised best by distinct conditional distributions. Specifically, each series can follow its own conditional distribution while the joint conditional distribution is captured with a copula.

The indicator proposed in this chapter aims at capturing cross-border movements, which in turn can be used as an indicator of migration net flow. The resulting indicator might differ from typical migration statistics because it is not possible to measure the intended durability of these movements. But this need not be a limitation; a timely approximation of migration net flow during the pandemic for instance, where intentions to settle or leave are in any case very uncertain, remains valuable to policy makers. The migration indicator based on Google newspaper queries will serve as an alternative high-frequency and fast indicator of migration alongside other statistics, as well as a covariate for interpolating short-term movements in other migration statistics.

The next section discusses the literature using big data for measuring human mobility and how this chapter contributes to it. Section 6.3 presents the modelling approach and the innovative feature relying on copulae. In Section 6.4 the performance of the model in estimating migration net flows is evaluated using New Zealanders migration in Australia. Section 6.5 concludes.

6.2 Comparison with existing studies on human mobility using big data

There is now a large body of literature focusing on using new internet data sources and big data in general to answer economic and social questions, starting with the seminal work of Choi and Varian (2012) showcasing the benefit of Google Trends. The last decade in particular has seen an increasing interest in exploiting big data to overcome the limitations of surveys, censuses, and administrative data in measuring human mobility and migration.

The International Organization for Migration and the European Commission held a workshop in 2017 aimed at reviewing and fostering advancement in this topic; IOM (2018) provides a summary. Big data sources potentially helpful in measuring human mobility can be divided broadly into three categories (IOM (2018)): mobile phone call records, geolocated social media data, internet activity (mainly google searches), and IP addresses from log-ins and emails.

Geolocalisation data stemming from mobile phone call records provide very timely and granular information on human mobility. They can notably be exploited for understanding migration flows within countries (Blumenstock (2012)), general patterns in human mobility (Gonzalez et al. (2009)) and forced displacements following specific events such as natural disasters (Lu et al. (2016)).

The highly detailed local analysis of human mobility possible with mobile phone call records is not feasible with the Google Trends-based indicator proposed in this chapter. Instead, the aim of the proposed indicator is to track human cross-border movements on a large scale, that is at country level. Moreover, unlike mobile phone records, Google Trends data are freely available and global. Consequently, once the Google Trends-based model is applicable to one country it is applicable to any country; the only difficulty is to find the name of the popular national newspapers.

Separately, the use of geolocated social media data, logins and emails for measuring global trends in human mobility has produced promising results. Zagheni et al. (2017) use Facebook's advertising platform for measuring the stock of migrants in a selection of OECD countries. State et al. (2014) analyse yearly trends in migration of workers in the US using LinkedIn data. Hawelka et al. (2014) investigate global mobility patterns using tweets sent in 2012. State et al. (2013) use geographic localisations from Yahoo! logins and manage to differentiate short and long-terms cross-border movements. Zagheni et al. (2017) use geographic locations mapped from IP addresses attached to Yahoo! email messages and provide a detailed analysis of migration rates from 2009 to 2011.

The approach based on Google newspaper queries suggested in this chapter complements the research mentioned above in three ways. First, Google is probably the most used online service (Alexa.com, which ranks the most popular websites in each country, ranks Google first in most of them), limiting the selection bias present when using less popular online services. Second, data gathered from Facebook and LinkedIn lag users' movements because a large share of users do not update their profiles in a timely way.

Therefore, while these services can be useful for approximating stocks of migrants, they are less useful for estimating net flows. Third, most studies are significantly limited in time, typically a few years and sometimes less, rendering a comparison with official migration data (generally released yearly) difficult. In contrast, the empirical application in this chapter benchmarks the performance of the google-based migration net flow indicator using official migration data over more than ten years.

The use of Google Trends has already been investigated for measuring human mobility. Kostakos et al. (2018) and Connor (2017) use Google Trends for improving predictions of refugees' arrival in Europe over a two-year period. Kostakos et al. (2018) survey the most common keywords searches by refugees and subsequently investigate the correlation between Google searches corresponding to these keywords with migrants' arrivals in specific locations. Using many keywords allows them to carry out a weekly and refined geographic analysis. However, the greater the number of key words used, the greater the number of factors which can drive the variation in the data. Although this risk is mitigated by their use of migrants' related languages such as Arabic and Farsi, it makes their approach difficult to apply to longer periods of time. Using a different approach, Connor (2017) analyses only two keywords searched in Turkey, "German" and "Greece" in Arabic, to predict the arrivals of migrants in Europe. The use of Google Trends provides an insight into refugees' arrivals but the empirical analysis is limited.

Focusing on migration trends more broadly, Wanner (2020), Bohme et al. (2020) and Wladyka (2017) use keywords related to migration for predicting short-term and real-time migration flows. Their approach consists of capturing the intention to migrate using economic and migration-related searches on Google and using it as a predictor for migration flows. This chapter adds to these studies by showing how keywords which capture migration flows directly, rather than the intention t migrate, can be used for real-time prediction. In doing so a comprehensive approach for tackling the conceptual and statistical challenges mentioned above, which are pervasive when using Google Trends, is proposed using an innovative time-series model.

6.3 A quasi score driven bivariate model with a copula

The indicator of cross-border movements based on Google search queries is defined as the diverging trend between two series: the series representing search queries in a given country and the series representing search queries everywhere. Therefore, filtering methods are necessary for separating the underlying trends in the data from, notably, seasonality and noise. Gaussian state space filtering techniques (see Harvey (1989) and Durbin and Koopman (2012)) are not suitable here because Google Trends data are often not (conditionally) normally distributed. This section presents a novel approach relying on score driven techniques and copulae for filtering jointly related series following distinct conditional distributions.

Score driven techniques, introduced by Creal et al. (2013) and Harvey (2013), provide an efficient and computationally light strategy to extract unobserved components in data subject to non-Gaussian features. Harvey and Luati (2014) discuss how time-varying trend and seasonal components can be modelled in a univariate score driven model with a Student's t-distribution. D'Innocenzo et al. (2021) extend this approach to a multivariate Student's t, but all series are constrained to have identical tail parameters; a limitation stressed by the authors. On the other hand, Creal et al. (2014) derive a multivariate model where each series can have its own conditional distribution, which introduces considerable flexibility, but estimation relies on the series being conditionally independent (no dependence in the prediction errors across series) which is not a feasible assumption here.

In this section a third class of multivariate model is proposed where each series can have a distinct conditional distribution (as in Creal et al. (2014)) but remains conditionally dependent to the other series (as in D'Innocenzo et al. (2021)). This is achieved by modelling the joint conditional distribution with a copula, producing a model which can feature a strong heterogeneity across series and capture complex cross-dependence patterns.

But unlike the models of Creal et al. (2014) and D'Innocenzo et al. (2021), the dynamic in the unobserved components does not come from the score of the joint conditional observation density. Instead, the dynamic in the model comes from the score of a constrained version of the observation density where the effect of the copula is alleviated by assuming independence. Therefore, while the copula is used to capture the dependence in the prediction errors and improve estimation, it does not affect the dynamics in the model; the score driven equation is thus disconnected from the log likelihood of the model. This is not necessarily a problem; in fact Blasques et al. (2020) derive a new class of score driven model where disconnecting the score driven mechanism from the observation density is the innovative feature.

6.3.1 A copula for modelling the joint conditional distribution

To retrieve an indicator of net migration of people from country A in country B, a bivariate model is applied to Google searches related to national newspapers of country A. Specifically, two series extracted from Google Trends are used for estimation: $y_{1,t}$, the log of Google queries originating country B, and $y_{2,t}$, the log of Google queries from everywhere. Each series can have a distinct conditional density such that

$$y_{i,t} = f_i(y_{i,t}|a_t;\Theta), \quad i = 1, 2, \quad t = 1, ..., N,$$

where a_t is a vector of unobserved components and Θ is a vector of fixed parameters. The joint conditional distribution of $y_{1,t}$ and $y_{2,t}$ is derived using a copula. Patton (2006) provides a conditional copula theory fitted for modelling the joint distribution of $y_t = (y_{1,t}, y_{2,t})'$ conditional on the past observations $y_{t-1}, ..., y_1$.

Copulae offer a great deal of flexibility in the modelling of dependencies. The form of the dependence is determined by the choice of the copula while its strength depends on the copula's parameters. They are notably popular in the actuarial and finance literature because they permit the modelling of systemic risk. The idea is that large and disruptive shocks are more likely to generate comovement than regular shocks. This reasoning can be applied to some macroeconomic variables which, despite showing little comovement in normal times, are affected similarly by an economic or financial crisis because of spill over effects. Google Trends series are another type of data which could exhibit such properties.

A copula $C(F_1(y_{1,t}), F_2(y_{2,t}))$ is a joint *cdf* where $F_1(.)$ and $F_2(.)$ are the marginal *cdf*'s. The corresponding joint *pdf* is

$$\begin{aligned} \frac{\partial^2 C(F_1(y_{1,t}), F_2(y_{2,t}))}{\partial y_{1,t} \partial y_{2,t}} &= \frac{\partial^2 C(F_1(y_{1,t}), F_2(y_{2,t}))}{\partial F_1(y_{1,t}) \partial F_2(y_{2,t})} \frac{\partial F_1(y_{1,t})}{\partial y_{1,t}} \frac{\partial F_2(y_{2,t})}{\partial y_{2,t}}, \\ &= c(F_1(y_{1,t}), F_2(y_{2,t})) f_1(y_{2,t}) f_2(y_{2,t}), \end{aligned} \tag{6.1}$$

where $c(F_1(.), F_2(.))$ is the *copula density* and $f_1(.)$ and $f_2(.)$ are the marginal *pdf*'s.

The conditional joint log-density for $y_t = (y_{1,t}, y_{2,t})'$ is

$$\begin{aligned} \text{log} f(y_t|a_t;\Theta) &= \text{log} c(F_1(y_{1,t}|a_t;\Theta), F_2(y_{2,t}|a_t;\Theta)|a_t) \\ &\quad + \text{log} f_1(y_{1,t}|a_t;\Theta) + \text{log} f_2(y_{2,t}|a_t;\Theta), \quad t = 1, ..., N. \end{aligned} \tag{6.2}$$

In the case of cross-sectional independence the copula density is unity and the joint log-density becomes simply the sum of each marginal log-density. The copula density distributions exploited in this chapter belong to elliptical and archimedean copulae, two common classes of copulae.

Elliptical copulae are constructed using elliptical distributions, typically the normal and student's t distributions.

Normal copula

$$C(u_1, u_2) = \Phi_2(\Phi^{-1}(u_1) + \Phi^{-1}(u_2)),$$

where Φ_2 is the cumulative distribution function of a bivariate normal with mean zero and covariance matrix equal to $R \in [-1, 1]^{2\times 2}$ with ones on the diagonal, and where Φ^{-1} is the quantile function of a standard normal distribution.

Student's t copula

$$C(u_1, u_2) = t_{2,\nu}(t_\nu^{-1}(u_1) + t_\nu^{-1}(u_2)),$$

where $t_{2,\nu}$ is the cumulative distribution function of a bivariate student's t with degrees of freedom set to ν, mean zero and covariance matrix equal to $R \in [-1, 1]^{2\times 2}$ with ones on the diagonal, and where t_ν^{-1} is the quantile function of a standard student's t with ν degrees of freedom.

Archimedean copulae, on the other hand, satisfy

$$C(u_1, u_2) = \phi^{-1}(\phi(u_1) + \phi(u_2))$$

where ϕ is called the generator. Four of the most common archimedean copulae are the Gumbel, Clayton, Frank and Joe copulae.

Gumbel copula

The Gumbel copula is given with the generator $\Phi(t) = (-\ln(t))^{\rho}$, where $\Theta \geq 1$, such that

$$C(u_1, u_2, \theta) = \exp\Big[- \Big((-\ln u_1)^{\theta} + (-\ln u_1)^{\theta} \Big)^{\frac{1}{\theta}} \Big].$$

The copula density is

$$\begin{aligned} c(u_1, u_2, \theta) &= \frac{\partial^2 C(F_1(y_{1,t}), F_2(y_{2,t}))}{\partial F_1(y_{1,t}) \partial F_2(y_{2,t})} \\ &= C(u_1, u_2, \theta) \frac{1}{u_1.u_2} ((-\ln u_1)^{\theta} + (-\ln u_2)^{\theta})^{-2+\frac{2}{\theta}} (\ln u_1 \ln u_2)^{\theta-1} \times \\ & (1 + (\Theta - 1)((-\ln u_1)^{\theta} + (-\ln u_2)^{\theta})^{\frac{1}{\theta}}). \end{aligned}$$

Clayton copula

The Clayton copula is given with the generator $\Phi(t) = (t^{-\Theta} - 1)/\theta$, where $\theta[-1, \infty)0$, such that The copula density is

$$c(u_1, u_2, \theta) = \frac{(1+\theta)(u_1.u_2)^{-(\theta+1)}}{(u_1^{-\theta} + u_2^{-\theta} - 1)^{\frac{1}{\theta}+2}}$$

Frank copula

The copula density is

$$c(u_1, u_2, \theta) = \theta(1 - e^{-\theta}) e^{-\theta(u_1+u_2)} [(1 - e^{-\theta}) - (1 - e^{-\theta u_1})(1 - e^{-\theta_2})]^{-2}$$

Joe copula

The copula density is

$$c(u_1, u_2, \theta) = ((1-u_1)^{\theta} + (1-u_2)^{\theta} - (1-u_1)^{\theta}(1-u_2)^{\theta})^{\frac{1}{\theta}-1} (1-u_1)^{\theta-1} (1-u_2)^{\theta-1}$$

Figure 6.2 illustrates the dependence structure arising from these six copula families plus the independent copula. For this, quantile values have been generated from bivariate copulae using the *copula* package in R (Yan (2007) and R Core Team (2021)). Any distribution could be applied to these set of quantiles to generate random observations.

Panel (a) illustrates the independent copula where both marginals are simply independent. Panel (b) illustrates the Gaussian copula with a correlation parameter of 0.5. Panel (c) shows the Student-t copula with dependence parameter and degrees of freedom set to 0.5 and 3. Panel (d) shows again the Student-t copula but with the copula parameter set to 0; this an interesting case because although the covariance is zero, both marginals are not independent. This is general feature of Student-t distribution when the degrees of freedom are not very large.

Panel (e), (f), (g) and (h) illustrate the Clayton, Gumbel, Frank and Joe copulae, which are four common Archimedean copulae. With the Clayton copula the dependence is asymmetric and stronger in the left side of the distributions. It is the opposite with the Joe copula; the dependence is stronger in the right side of the distribution. The Gumbel copula yields a greater dependence in the right tails, but the dependence also remains strong on the left side. The Frank copula, on the other hand, yields a symmetric dependence structure.

6.3.2 Skewed Student's t marginals

Each series' conditional density comes from the family of asymmetric student-t (AST) density of Zhu and Galbraith (2010). The distinctive features of the AST are its shape parameter, or skewness parameter, which controls the asymmetry round the central part of the distribution, and its tail parameters which control tail behaviour independently on each side. In this application the tail parameters are constrained to be identical on both sides for each series.

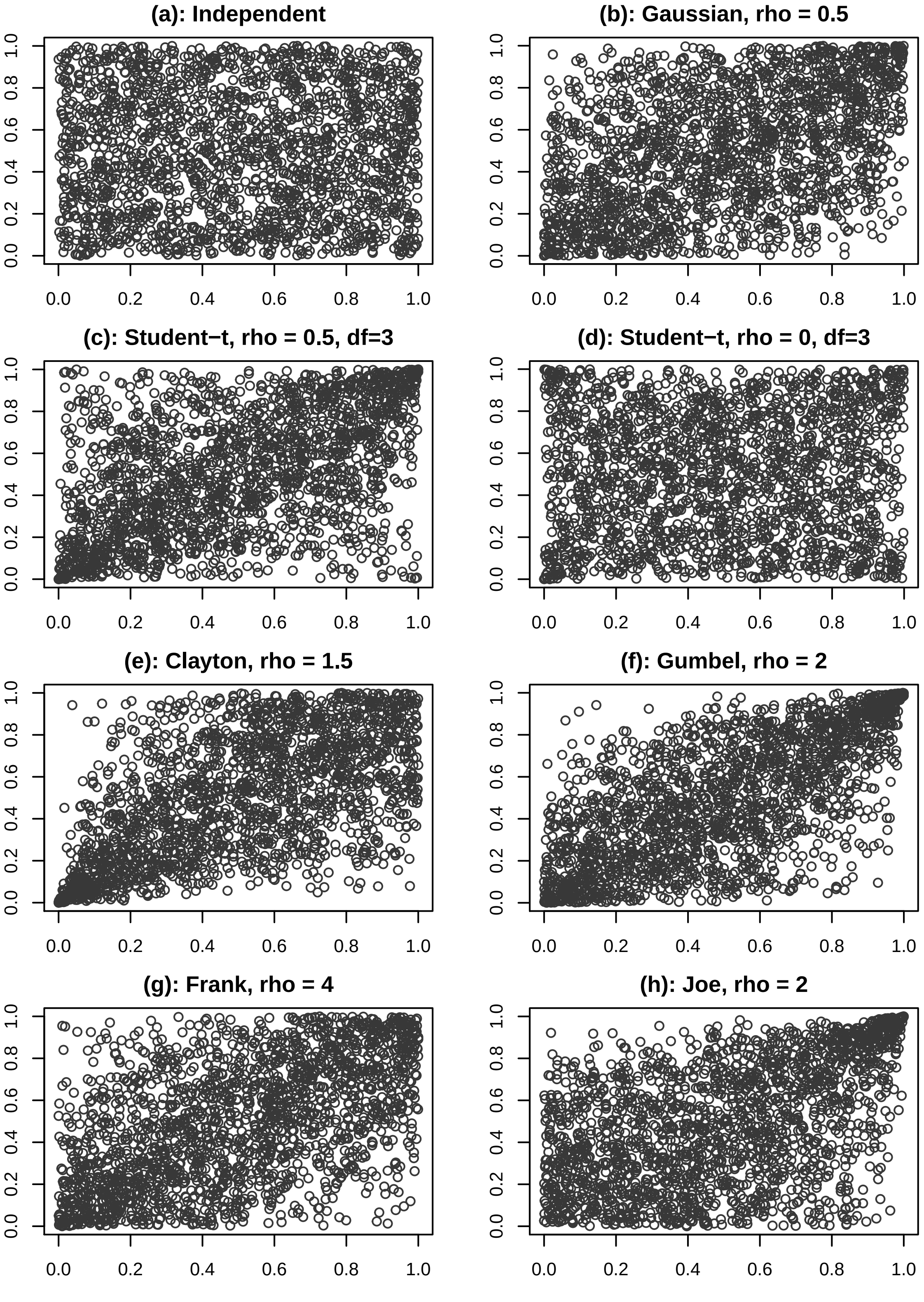


Figure 6.2: 2000 draws of uniforms. Comparison of the seven copulae fitted in the empirical application.

The log AST density of observation t of series i takes the form

$$\begin{aligned}\log f_i(y_{i,t}|Y_{t-1}) = &-\ln\sigma_{i,t} - \frac{\nu_{1,i}+1}{2}\ln\Big[1+\frac{1}{\nu_{1,i}}\Big(\frac{y_{i,t}-\mu_{i,t}}{2\alpha_{i,t}\sigma_{i,t}K(\nu_{1,i})}\Big)^2\Big]\ 1(y_{i,t}\leq\mu_{i,t})\\ &-\frac{\nu_{2,i}+1}{2}\ln\Big[1+\frac{1}{\nu_{2,i}}\Big(\frac{y_{i,t}-\mu_{i,t}}{2(1-\alpha_{i,t})\sigma_{i,t}K(\nu_{2,i})}\Big)^2\Big]\ 1(y_{i,t}>\mu_{i,t}),\end{aligned} \tag{6.3}$$

where $\sigma_{i,t}$ is the scale parameter, $\alpha_{i,t}$ is the shape parameter which can take values in $(0,1)$, and $\nu_{1,i}$ and $\nu_{2,i}$ are respectively the left and right tail parameters which take positive values. $K(\nu)=\Gamma((\nu+1)/2)/(\sqrt{\nu\pi}\Gamma(\nu/2))$ ($\Gamma(.)$ is the Gamma function) and $1(x)$ is an indicator variable equal to one if statement x is true and zero otherwise. The distribution is skewed towards positive values if $\alpha_{i,t}<0.5$ and towards negative values if $\alpha_{i,t}>0.5$. When the tail parameters are constrained to be very large and the skewness parameter to 0.5 the AST is equivalent to a (scaled) normal distribution.

6.3.3 A quasi score driven dynamic

One of the main advantage of the score driven models of Creal et al. (2013) and Harvey (2013) is their ability to handle easily a wide range of distributions and nonlinearities. Chapters 4 and 5 notably illustrate their use for nowcasting and cleaning. However, tying the driving force in the parameters' variations to the conditional distribution of the observations can be problematic. For instance, the score downweights the effect of large shocks in the dynamic parameters when the data are characterised best by a fat tailed distribution. While this is useful if these large shocks are one-off movements or measurement errors, it becomes limiting if they carry economic meaning and indicate structural breaks which should be reflected in the parameters of the model.

Blasques et al. (2020) suggest overcoming this limitation by allowing the conditional distribution behind the score to vary from the conditional distribution of the observations, thus creating *quasi* score driven models. A particularly convenient form of this new type of models arises from using a constrained version of the conditional distribution of the observation to derive the score. In the example mentioned above, the score can be taken with respect to a distribution where the tail parameters are constrained to be very large, thus forcing the score to transmit large variations in the data to the unobserved components or dynamic parameters.

Another benefit of quasi score driven models is their ability to reduce computational burdens and instabilities arising from the score when the conditional distribution of the observation introduces large nonlinearities; this is the case when using copulae. For many distributions the hessian used for scaling the score is also not easily calculable and evaluating it requires numerical methods. Using a simpler form of the conditional distribution to derive the score is a good way to avoid this issue.

This chapter adopts a quasi score driven approach by imposing cross-sectional independence in the conditional distribution behind the score. This means that the copula has no effect on the score.

The dynamic equation for the latent states is

$$a_{t+1} = Ta_t + Ks_t, \tag{6.4}$$

where s_t denotes the scaled first derivative of the log density with respect to the vector a_t:

$$s_t = S_t \Delta_t', \quad \Delta_t = \frac{\partial \text{log} f^*(y_t|Y_{t-1})}{\partial a_t}, \tag{6.5}$$

The matrix K is a diagonal matrix of gains which are estimated via maximum likelihood along the unknown elements of the matrix T and the initial vector a_1. The matrix S_t is a scaling matrix set to the Moore–Penrose inverse of the expected information matrix [1]

[1] The information matrix needs to be inverted at each step of the recursion because its components are dynamic which considerably slows down estimation. To overcome this technical issue the score driven recursion (hence the log likelihood evaluation as well) is written in C++ (I am grateful to Caterina Schiavoni for her help with this) while the optimisation and the remainder of the analysis is

given by

$$\mathcal{I}_t = \mathrm{E}\big[\Delta_t'\Delta_t|Y_{t-1}\big]. \tag{6.6}$$

Since the log density (5.2) relies on conditional independence, the score can be expressed conveniently as

$$\Delta_t = \sum_{i=1}^{2} \delta_{i,t}\Delta_{i,t}, \tag{6.7}$$

where $\Delta_{i,t}$ is the score of series i, while the expected information matrix is

$$\mathcal{I}_t = \sum_{i=1}^{2} \delta_{i,t}\mathcal{I}_{i,t}. \tag{6.8}$$

Using the chain rule yields

$$\Delta_{i,t} = \frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial a_t} = \frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial \mu_{i,t}} . \frac{\partial \mu_{i,t}}{\partial a_t} \tag{6.9}$$

while the expected information matrix becomes

$$\mathcal{I}_{i,t} = \Big(\frac{\partial \mu_{i,t}}{\partial a_t}\Big)' \mathrm{E}\Bigg[\bigg(\frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial \mu_{i,t}}\bigg)^2 |Y_{t-1}\Bigg] \frac{\partial \mu_{i,t}}{\partial a_t}. \tag{6.10}$$

The score and information element for location parameters are shown in appendix E.

The next section details the dynamics of the components in a_t.

6.3.4 A bivariate unobserved components model with a common trend

Both series are assumed to follow a bivariate structural model given by

$$\begin{aligned} y_{1,t} &= \mu_{1,t} + \Phi_\mu \mu_{2,t} + \gamma_{1,j,t} + v_{1,t}, \\ y_{2,t} &= \mu_{2,t} + \gamma_{2,j,t} + v_{2,t}. \end{aligned} \tag{6.11}$$

carried in R.

The time-varying trends in the data are captured with $\mu_{1,t}$ and $\mu_{2,t}$, the seasonal effects by $\gamma_{1,j,t}$ and $\gamma_{2,j,t}$, $j = 1, ..., 12$, while $v_{1,t}$ and $v_{2,t}$ are the prediction errors. The trend $\mu_{2,t}$ represents the general trend in newspaper reading and affects both the country-specific and global series. The trend $\mu_{2,t}$, on the other hand, affects only the country-specific series and should be governed by the net flow of people from A in the region.

Trend and seasonal components follow random walk models given by

$$\begin{aligned} \mu_{i,t+1} &= \mu_{i,t} + K_i^{\mu} s_{i,t}^{\mu}, \\ \gamma_{i,j,t+1} &= \gamma_{i,j,t} + K_i^{\gamma} s_{i,t}^{\gamma}, \qquad j = 1, ..., 12, \end{aligned} \tag{6.12}$$

$i = 1, 2$, where the sum of the seasonal variations is constrained to be zero across seasonal effects in each period following the method of Harvey and Luati (2014).

6.3.5 Maximum likelihood estimation

The vector of unknown parameters which includes the marginals' parameters, the copula dependence parameter (plus the degree of freedom for the student-t copula) and the initial values of the time-varying parameters a_1 are estimated via Maximum Likelihood as

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{t=1}^{N} \log f(y_t|Y_{t-1}), \tag{6.13}$$

where $\log f(y_t|Y_{t-1})$ is the is given by (6.2). As in previous chapters, the minimisation of the negative log likelihood function is carried out with the BFGS algorithm in R. To improve estimation an analogue unobserved component state space model is estimated and the smoothed state vector is used to initialise the estimation of the score driven model. This is necessary to get good estimates of seasonal components.

6.3.6 Monte Carlo simulations

In this section, the model presented above is estimated repetitively on simulated data to investigate the feasibility of estimation and its potential biases.

The Monte Carlo experiment takes the following form. First, a bivariate vector of observation errors $v_t = (v_{1,t}, v_{2,t})'$, $t = 1, ..., N$, following a copula distribution with distinct skewed student-t margins is drawn using the R package copula (Yan (2007)). The copula used for the experiment is the Gumbel copula, but using different copulae does not alter the results. Second, these observations errors are used to retrieve the vector of observations $y_t = (y_{1,t}, y_{2,t})'$ using (6.11). Third, the model is estimated via Maximum Likelihood using the thus generated observations. For simplicity the seasonal terms are set to zero; this should not affect the conclusions which can be drawn from the experiments.

The Google Trends data used for estimating the model in the empirical application below contain round 165 observations. Thus to understand the implication of using a relatively small sample the simulation are drawn with $N = 165$, $N = 500$ and $N = 1000$. Table 6.1 and 6.2 show the results when the copula parameter is set to 2 and 4 receptively. They show the mean estimates of the unknown parameters with the associated bias and root-mean-square-error calculated as

$$\text{Bias}(\hat{\sigma}_1) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_{1,i} - \sigma_1), \qquad \text{RMSE}(\hat{\sigma}_1) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_{1,i} - \sigma_1)^2},$$

where $\hat{\sigma}_{1,i}$ is the Maximum Likelihood estimate of σ_1 from simulation i.

All parameters except the degrees of freedom attached to second variable, ν_2, are estimated precisely starting from 165 observations. The bias and RMSE for ν_2, which has a true value of 7, are larger than for ν_1, which is equal to 3. This suggests that the larger the tail parameters, the less precisely they are estimable. This is because the effect

Table 6.1: Results of the Monte Carlo experiments. Copula parameter $\theta = 2$.

		N = 165			N = 500			N = 1000		
	True value	Mean	Bias	RMSE	Mean	Bias	RMSE	Mean	Bias	RMSE
σ_1	0.05	0.050	-0.0002	0.004	0.050	-0.0001	0.002	0.050	-0.0001	0.002
σ_2	0.1	0.098	-0.002	0.008	0.099	-0.001	0.004	0.100	-0.0003	0.003
ν_1	7	10.126	3.126	15.844	7.359	0.360	1.771	7.189	0.190	1.162
ν_2	3	3.051	0.051	0.642	3.008	0.008	0.278	3.003	0.003	0.189
α_1	0.6	0.601	0.001	0.038	0.600	0.0001	0.020	0.600	-0.0001	0.013
α_2	0.4	0.396	-0.004	0.029	0.399	-0.001	0.013	0.400	-0.0003	0.009
θ	2	2.024	0.024	0.182	2.015	0.015	0.089	2.003	0.004	0.059
k_1	0.3	0.298	-0.002	0.045	0.300	0.0004	0.022	0.300	-0.0002	0.016
k_2	0.7	0.711	0.011	0.072	0.703	0.003	0.034	0.702	0.002	0.022
a_1	3	3.005	0.005	0.038	3.005	0.005	0.036	3.003	0.003	0.036
a_2	3	3.025	0.025	0.135	3.023	0.023	0.129	3.026	0.026	0.134

Note: Results from 1000 simulations. N indicates the number of observations in each series. Estimation with the Gumbel copula.

Table 6.2: Results of the Monte Carlo experiments. Copula parameter $\theta = 4$.

		N = 165			N = 500			N = 1000		
	True value	Mean	Bias	RMSE	Mean	Bias	RMSE	Mean	Bias	RMSE
σ_1	0.05	0.050	-0.0001	0.003	0.050	0	0.002	0.050	-0.0001	0.001
σ_2	0.1	0.098	-0.002	0.007	0.100	-0.0005	0.005	0.100	-0.0005	0.003
ν_1	7	7.630	0.630	2.856	7.209	0.209	1.143	7.028	0.028	0.648
ν_2	3	3.014	0.014	0.373	3.022	0.022	0.413	2.999	-0.001	0.140
α_1	0.6	0.604	0.004	0.031	0.602	0.002	0.020	0.602	0.002	0.017
α_2	0.4	0.394	-0.006	0.028	0.398	-0.002	0.016	0.399	-0.001	0.012
θ	4	4.035	0.035	0.570	3.995	-0.005	0.337	3.989	-0.012	0.280
k_1	0.3	0.296	-0.004	0.035	0.298	-0.002	0.019	0.298	-0.002	0.015
k_2	0.7	0.713	0.013	0.072	0.705	0.005	0.043	0.703	0.003	0.033
a_1	3	3.003	0.003	0.029	3.004	0.004	0.031	3.005	0.005	0.033
a_2	3	3.022	0.022	0.112	3.028	0.028	0.134	3.032	0.032	0.127

Note: Results from 1000 simulations. N indicates the number of observations in each series. Estimation with the Gumbel copula.

that marginal variations in the degrees of freedom have on the log likelihood decreases as the degrees of freedom increase. Overall this Monte Carlo experiment proves the feasibility of estimation of this new class of quasi score driven model relying copulae even with a limited number of observations; the model can be confidently applied to the Google Trends data.

6.4 Application with New Zealanders net migration in Australia

The idea exposed in this chapter consists of using the divergence in popularity of newspaper queries on Google in a given country compared to all regions as an indicator of migration net flow. To test this proposition and evaluate the performance of the indicator based on Google newspaper queries to track official migration statistics, this section applies the model to migration data from Australian Bureau of Statistics and newspapers from New Zealand. Migration data from Australia are renowned for their quality and the flow of New Zealanders in Australia is significant.

Figure 6.1 above shows the data used for estimation. The Google Trends series are retrieved with the *gtrendsR* package in *R* (Massicotte and Eddelbuettel (2021)) using the keywords "nz herald + stuff nz + otago daily + scoop nz" which correspond to four popular newspapers in New Zealand.

The first step consists of choosing the appropriate copula family. The model is estimated with each copula, i.e. independent, Clayton, Gumbel, Frank or Student's t copula (which encompasses the normal copula when the tail parameter is large). The multivariate Student-t model of D'Innocenzo et al. (2021) is also estimated to investigate the advantage of estimating distinct tail and shape parameters for each series.

Table 6.3 shows the estimated parameters and the maximum log likelihood found for each model. Each copula model gives a similar picture: the tail parameters are relatively large and both series are skewed towards positive values, with a significantly more pronounced skewness for the second series. The model chosen to retrieve the indicator of New Zealanders net flow in Australia is the Gumbel copula model on the grounds that it yields the highest log likelihood.

Table 6.3: Estimation with Google newspaper queries from New Zealand. Each model has a distinct joint conditional distribution which include 6 copulae (counting the independent copula), a multivariate Student-t and the Normal distribution.

Model	$\hat{\sigma_1}$	$\hat{\nu_1}$	$\hat{\alpha_1}$	$\hat{\sigma_2}$	$\hat{\nu_2}$	$\hat{\alpha_2}$	$\hat{\rho}$	$\hat{\nu_{cop}}$	$\hat{k}_1$	$\hat{k}_2$	$\hat{\Phi_\mu}$	Log. Lik
Indep.	0.115	1.987	0.612	0.081	3.642	0.269	0		0.368	0.368	1.221	218.732
Clayton	0.118	2.033	0.628	0.087	4.262	0.306	0.597		0.364	0.364	1.24	233.455
Gumbel	0.134	2.531	0.589	0.082	3.779	0.3	1.52		0.326	0.326	1.212	246.74
Frank	0.117	2.102	0.617	0.081	3.875	0.272	0.082		0.39	0.39	1.22	219.886
Joe	0.129	2.289	0.564	0.08	3.828	0.286	1.755		0.34	0.34	1.201	245.07
St t cop	0.133	2.577	0.602	0.084	4.181	0.316	0.52	3.383	0.335	0.335	1.2	**248.302**
St t	0.137	2.513	0.5	0.062	2.513	0.5	0.043		0.734	0.734	0.075	239.015
Normal	0.232	100	0.5	0.099	100	0.5	0.049		0.251	0.251	0.021	186.894

Figure 6.3 shows a scatter plot of the quantiles associated to the prediction errors derived with the Student-t copula model. These are derived simply by applying the cumulative distribution function of the asymmetric Student-t distribution on the prediction errors. The dependence is stronger on the right tails which explains why the Gumbel, Joe and Student-t copula yield significantly greater log likelihoods than the other models. The better fit of the Student-t copula stems from the occurrence of observations in opposite tails, something unlikely with the archimedean families used here and the Gaussian copula.

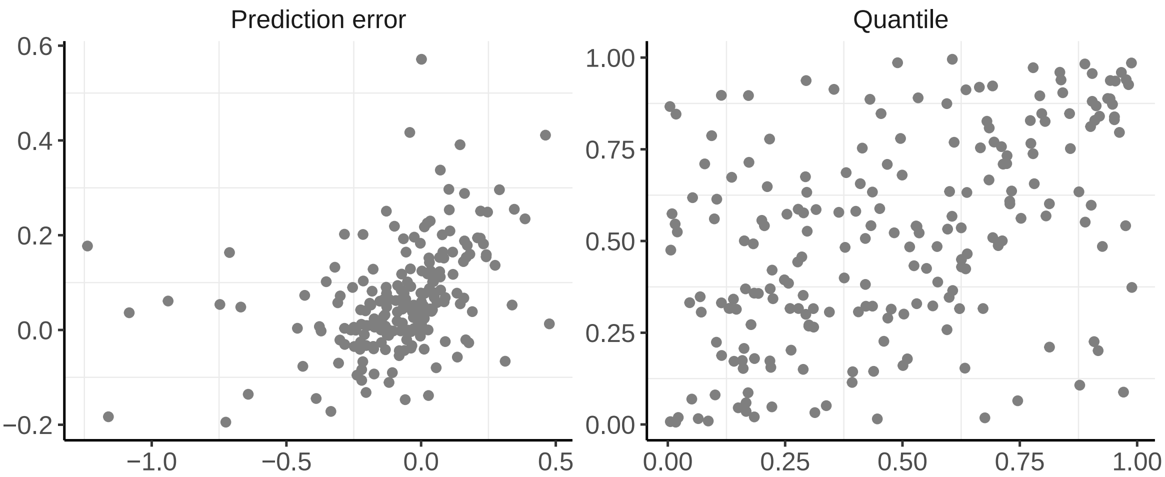


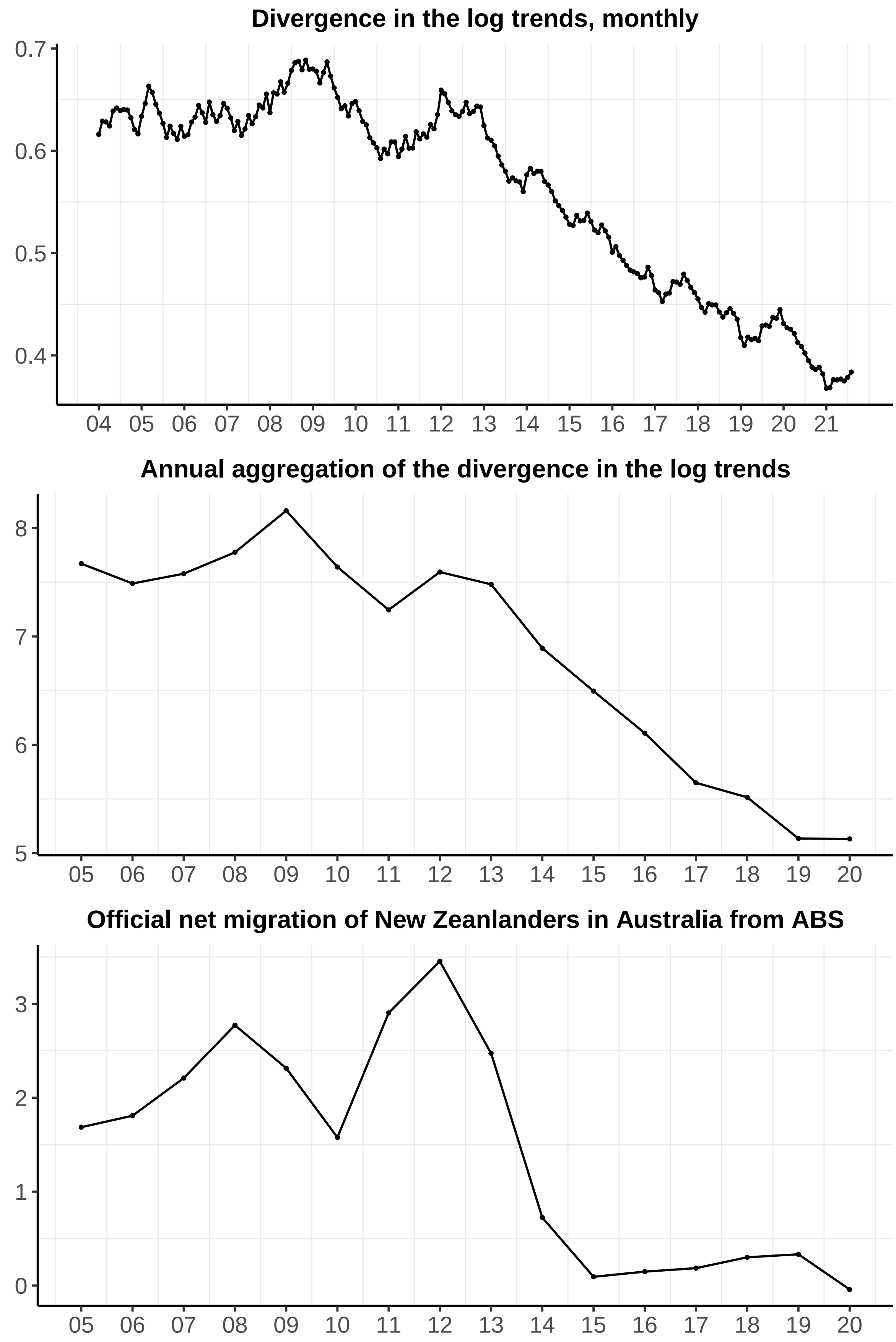
Figure 6.3: Predictions errors and associated quantiles from the estimation of the Student's copula model with the Google Trends data on newspaper queries.

The first plot of figure 6.4 shows the estimated diverging trend between the Australian and world series using the model featuring the Gumbel copula. This trend represents the indicator of New Zealanders net flow in Australia. It is a monthly indicator but for comparison with the ABS migration data, which are yearly (using financial years), the second plot aggregates the monthly estimates to annual figures representing financial years. The annual figures tend to follow closely the ABS data which are shown in the third plot. There is a decrease in 2010 followed by a general increase until 2012. From 2012 onward there is a constant decrease until 2015 in the ABS data which appears as well in the Google Trends indicator, but while the ABS data tends to stagnate after that, the Google Trends indicator keeps falling, albeit at a lower rate.

Having established that the Google-based indicator tracks broadly the official migration data, it is possible to derive from it a timely insight on the pandemic period. The monthly indicator based on newspaper queries decreases sharply in the second half of 2020, suggesting that New Zealanders have returned home as the emergence of a long-lasting global pandemic became apparent.

6.5 Conclusion

This chapter shows that newspaper queries extracted from Google Trends can be used to derive a timely indicator of migration net flow. For this a flexible score driven model relying on copulae is proposed where the innovative feature lies in its ability to model jointly series which are related but characterised best by distinct conditional distributions. The case study with New Zealanders net flow in Australia shows that the proposed indicator tracks official migration figures. The advantages of the indicator derived from Google Trends are its frequency and timeliness. Notably, it indicates an outflow of New Zealanders from Australia following the Covid-19 pandemic which does not yet appears in official data.


Divergence in the log trends, monthly

Annual aggregation of the divergence in the log trends

Official net migration of New Zeanlanders in Australia from ABS

Figure 6.4: Estimates derived from estimating the model with Gumbel copula using Google queries of national newspapers from New Zealand. This indicator is retrieved as $\exp(\hat{\mu}_{1,t})$. Data extracted with the with keyword ”nz herald + stuff nz + otago daily + scoop nz”.

Chapter 7

Conclusion

The chapters exploring the use of VAT returns in the UK to compute monthly GDP have shown the potential of state space methods for temporal disaggregation when the data are noisy. The extreme observation errors in the early vintages of the VAT, however, are treated best with a first-stage score driven cleaning approach downweighting outliers. Score driven techniques are the common theme of the last two chapters, where they are illustrated in different contexts. Chapter 5 shows the potential of score driven models for density nowcasting by introducing dynamic shape parameters. Tackling a different problem, chapter 6 proposes a new class of models relying on copulae to accommodate heterogeneous series. The proposed methodology relying on copulae for modelling unobserved components in highly heterogeneous series opens the door to a wide range of applications.

Bibliography

Adrian, T., N. Boyarchenko, and D. Giannone (2019, April). Vulnerable growth. *American Economic Review 109*(4), 1263–1289.

Antolin-Diaz, J., T. Drechsel, and I. Petrella (2017, May). Tracking the slowdown in long-run gdp growth. *The Review of Economics and Statistics 99*(2), 343–356.

Antolin-Diaz, J., T. Drechsel, and I. Petrella (2020). Advances in Nowcasting Economic Activity:Secular Trends, Large Shocks and New Data.

Aruoba, S. B., F. X. Diebold, and C. Scotti (2009). Real-time measurement of business conditions. *Journal of Business & Economic Statistics 27*(4), 417–427.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 367–389.

Banbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013). Now-Casting and the Real-Time Data Flow. In G. Elliott, C. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2 of *Handbook of Economic Forecasting*, Chapter 0, pp. 195–237. Elsevier.

Bean, C. (2016). Independent review of uk economic statistics. *https://www.gov.uk/government/publications/independent-review-of-uk-economicstatistics-final-report*.

Bell, W. R. and S. C. Hillmer (1983). Modeling time series with calendar variation. *Journal of the American Statistical Association 78*(383), 526–534.

Blasques, F., C. Francq, and S. Laurent (2020, October). A new class of robust observation-driven models. *20*(20-073/III).

Blasques, F., S. Koopman, M. Mallee, and Z. Zhang (2016). Weighted maximum likelihood for dynamic factor analysis and forecasting with mixed frequency data. *Journal of Econometrics 193*(2), 405–417.

Blumenstock, J. E. (2012). Inferring patterns of internal migration from mobile phone call records: evidence from rwanda. *Information Technology for Development 18*(2), 107–125.

Bohme, M. H., A. Griger, and T. Stohr (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics 142*, 102347.

Boot, J. C. G., W. Feibes, and J. H. C. Lisman (1967). Further methods of derivation of quarterly figures from annual data. *Applied Statistics 16*(1), 65.

Bowman, K. O. and L. R. Shenton (1975). Omnibus test contours for departures from normality based on vb 1 and b 2. *Biometrika 62*(2), 243.

Buccheri, G., G. Bormetti, F. Corsi, and F. Lillo (2021). Filtering and smoothing with score-driven models.

Caivano, M., A. Harvey, and A. Luati (2016). Robust time series models with trend and seasonal components. *SERIEs 7*(1), 99–120.

Carriero, A., T. E. Clark, and M. Marcellino (2016). Common drifting volatility in large bayesian vars. *Journal of Business & Economic Statistics 34*(3), 375–390.

Carriero, A., T. E. Clark, and M. Marcellino (2018). Measuring uncertainty and its impact on the economy. *The Review of Economics and Statistics 100*(5), 799–815.

Carriero, A., T. E. Clark, and M. Marcellino (2020). Assessing international commonality in macroeconomic uncertainty and its effects. *Journal of Applied Econometrics 35*(3), 273–293.

Choi, H. and H. Varian (2012). Predicting the present with google trends. *Economic Record 88*, 2–9.

Chow, G. C. and A.-L. Lin (1971). Best linear unbiased estimation of missing observations in an economic time series. *Review of Economics and Statistics 53*, 372–5.

Connor, P. (2017). The digital footprint of europe's refugees. *Pew Research Center*.

Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics 28*(5), 777–795.

Creal, D., B. Schwaab, S. J. Koopman, and A. Lucas (2014). Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics 96*(5), 898–915.

Delle-Monache, D., A. De-Polis, and I. Petrella (2020). Modelling and Forecasting Macroeconomic Downside Risk. *Economic Modelling and Forecasting Group* (34).

Delle-Monache, D. and I. Petrella (2017). Adaptive models and heavy tails with an application to inflation forecasting. *International Journal of Forecasting 33*(2), 482–501.

D'Innocenzo, E., A. Luati, and M. Mazzocchi (2021). A robust score-driven filter for multivariate time series.

Doz, C., L. Ferrara, and P.-A. Pionnier (2020, January). Business cycle dynamics after the great recession: An extended markov-switching dynamic factor model. (2020/01).

Durbin, J. and S. J. Koopman (2012). *Time series analysis by state space methods*. Oxford University Press.

Fernández, C. and M. F. Steel (1996). On bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association 93*(441), 359–371.

Fernandez, R. B. (1981). A methodological note on the estimation of time series. *The Review of Economics and Statistics 63*(3), 471–476.

Gatz, D. F. and L. Smith (1995). The standard error of a weighted mean concentration i. bootstrapping vs other methods. *Atmospheric Environment 29*(11), 1185–1193.

Gomez, H., F. Torres, and H. Bolfarine (2007). Large-sample inference for the epsilon-skew-t distribution. *Communications in Statistics - Theory and Methods 36*(1), 73–81.

Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi (2009). Understanding individual human mobility patterns. *Nature 458*(7235), 238–238.

Gorgi, P., S. J. Koopman, and M. Li (2019). Forecasting economic time series using score-driven dynamic models with mixed-data sampling. *International Journal of Forecasting 35*(4), 1735–1747.

Grassi, S., T. Proietti, C. Frale, M. Marcellino, and G. Mazzi (2015). Euromind-c: A disaggregate monthly indicator of economic activity for the euro area and member countries. *International Journal of Forecasting 31*(3), 712–738.

Hand, D. J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 181*(3), 555–605.

Harvey, A. and A. Luati (2014). Filtering with heavy tails. *Journal of the American Statistical Association 109*(507), 1112–1122.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Harvey, A. C. (2006). Seasonality and unobserved components models: an overview. *Conference on seasonality, seasonal adjustment and their implications for short-term analysis and forecasting*.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series (Econometric Society Monographs)*. Cambridge University Press.

Harvey, A. C. and S. J. Koopman (1997). Multivariate structural time series models. *System Dynamics in Economic and Financial Models, eds C. Heij, J. M. Shumacher, B. Hanzon and C. Praagman, Chichester: John Wiley & Sons Ltd*, 269–298.

Harvey, A. C. and R. G. Pierse (1984). Estimating missing observations in economic time series. *Journal of the American Statistical Association 79*(385), 125–131.

Hawelka, B., I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti (2014). Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science 41*(3), 260–271.

Helske, J. (2017). Kfas: Exponential family state space models in r. *Journal of Statistical Software 78*(10), 1–39.

Huber, F. (2016). Density forecasting using bayesian global vector autoregressions with stochastic volatility. *International Journal of Forecasting 32*(3), 818–837.

IOM (2018). Big data and migration, how data innovation can serve migration policy-making. *Data Bulletin*.

Koopman, S. J., A. Lucas, and M. Scharth (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *Review of Economics and Statistics 98*(1), 97–110.

Kostakos, P., A. Pandya, M. Oussalah, S. Hosio, A. Sattari, V. Kostakos, N. van Berkel, C. Breidbach, and O. Kyriakouli (2018). Correlating refugee border crossings with internet search data. *2018 IEEE International Conference on Information Reuse and Integration for Data Science*.

Litterman, R. B. (1983). A random walk, markov model for the distribution of time series. *Journal of Business & Economic Statistics 1*(2), 169–173.

Lu, X., D. J. Wrathall, P. R. Sundsoy, M. Nadiruzzaman, E. Wetter, A. Iqbal, T. Qureshi, A. Tatem, G. Canright, K. Enge-Monsen, and B. Linus (2016). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in bangladesh. *Global Environmental Change 38*, 1–7.

Lucas, A. and Z. Zhang (2016). Score-driven exponentially weighted moving averages and value-at-risk forecasting. *International Journal of Forecasting 32*(2), 293–302.

Mariano, R. S. and Y. Murasawa (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics 18*(4), 427–443.

Massicotte, P. and D. Eddelbuettel (2021). *gtrendsR: Perform and Display Google Trends Queries.* R package version 1.4.8.

McConnell, M. M. and G. Perez-Quiros (2000). Output fluctuations in the united states: What has changed since the early 1980's? *American Economic Review 90*(5), 1464–1476.

Mitchell, J., R. J. Smith, M. R. Weale, S. Wright, and E. L. Salazar (2005). An indicator of monthly gdp and an early estimate of quarterly gdp growth. *The Economic Journal 115*(501), 108–129.

Moauro, F. and G. Savio (2005). Temporal disaggregation using multivariate structural time series models. *The Econometrics Journal 8*(2), 214–234.

Patton, A. J. (2006). Modelling asymmetric exchange rate dependence. *International Economic Review 47*(2), 527–556.

Pettenuzzo, D., A. Timmermann, and R. Valkanov (2016). A midas approach to modeling first and second moment dynamics. *Journal of Econometrics 193*(2), 315–334.

Proietti, T. (2000). Comparing seasonal components for structural time series models. *International Journal of Forecasting 16*(2), 247–260.

Proietti, T. (2006). On the estimation of nonlinearly aggregated mixed models. *Journal of Computational and Graphical Statistics 15*(1), 18–38.

Proietti, T. and F. Moauro (2006). Dynamic factor analysis with non-linear temporal aggregation constraints. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 55*(2), 281–300.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Russ, B. and A. Tariq (2017). The impact of moving holidays on official statistics time series. *Proceedings of the 22nd GSS Methodology Symposium 2017*.

Salazar, E., R. Smith, M. Weale, and S. Wright (1997). A monthly indicator of gdp. *National Institute Economic Review 161*(1), 84–89.

Spiegelhalter, D. (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application 4*(1), 31–60.

State, B., M. Rodriguez, D. Helbing, and E. Zagheni (2014). *Migration of Professionals to the U.S.*, pp. 531–543. Cham: Springer International Publishing.

State, B., I. Weber, and E. Zagheni (2013). Studying inter-national mobility through ip geolocation. *WSDM'13, February 4–8, 2013, Rome, Italy*.

Stephens, M. and J. Allcoat (2016). Exploitation of hmrc vat data.

Wanner, P. (2020). How well can we estimate immigration trends using google data? *Quality & Quantity 55*(4), 1181–1202.

Wladyka, D. (2017). Queries to google search as predictors of migration flows from latin america to spain. *Journal of Population and Social Studies 25*(4), 312–327.

Yan, J. (2007). Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 555–605.

Zagheni, E., I. Weber, and K. Gummadi (2017). Leveraging facebook's advertising platform to monitor stocks of migrants. *Population and Development Review 43*(4), 721–734.

Zhu, D. and J. W. Galbraith (2010). A generalized asymmetric student- distribution with application to financial econometrics. *Journal of Econometrics 157*(2), 297–305.

Appendix A

Approximating nonlinear temporal aggregation constraints

A.1 Variables in log levels

A quarterly variable in levels Q_t must equal the three-month sum of its monthly subcomponents M_t:

$$Q_t = M_t + M_{t-1} + M_{t-2}. \tag{A.1}$$

To account for heteroskedasticity and multiplicative components, however, the data are generally taken in logarithms, such that the variables modelled are $q_t = \log(Q_t)$, and $m_t = \log(M_t)$. Since the sum of the logarithms is not equal to the logarithm of the sum, the accounting constraint takes a nonlinear form:

$$q_t = \log\Big[\exp(m_t) + \exp(m_{t-1}) + \exp(m_{t-2})\Big]. \tag{A.2}$$

It is generally not practical to work from nonlinear observation equations. The Kalman filter, notably, is not applicable in such cases. It is possible, however, to work from a very precise linear approximation due to Salazar et al. (1997) and Mitchell et al.

(2005). They derive an approximation for a given nonlinear function h(.) using a first order taylor expansion at the monthly average $\bar{m}_t = \frac{1}{3}\sum_{i=0}^{2} m_{t-i}$:

$$\text{h}(m_t) \approx \text{h}(\bar{m}_t) + \text{h'}(\bar{m}_t)(m_t - \bar{m}_t). \tag{A.3}$$

The quarterly sum of the monthly variable in levels is then

$$\begin{aligned}\sum_{i=0}^{2} \exp(m_{t-i}) \approx\ & 3\text{h}(\bar{m}_t) + \text{h'}(\bar{m}_t)(m_t - \bar{m}_t) \\ & + \text{h'}(\bar{m}_t)(m_{t-1} - \bar{m}_t) + \text{h'}(\bar{m}_t)(m_{t-2} - \bar{m}_t) \\ & \approx 3\text{h}(\bar{m}_t) + \text{h'}(\bar{m}_t)(\sum_{i=0}^{2} \bar{m}_{t-i} - 3\bar{m}_t) \\ & \approx 3\text{h}(\bar{m}_t) + \text{h'}(\bar{m}_t)(\sum_{i=0}^{2} \bar{m}_{t-i} - 3 \times \frac{1}{3}\sum_{m=0}^{3} m_{t-i}) \\ & \approx 3\text{h}(\bar{m}_t).\end{aligned} \tag{A.4}$$

Using the exponential function and plugging this approximation into the temporal aggregation constraint (A.2) gives

$$\begin{aligned} q_t &= \log\Big[3\text{exp}(\bar{m}_t)\Big], \\ &= \log\Big[3.\text{exp}(\frac{1}{3}\sum_{m=0}^{2} m_{t-i})\Big], \\ &= \log(3) + \frac{1}{3}m_t + \frac{1}{3}m_{t-1} + \frac{1}{3}m_{t-2}. \end{aligned} \tag{A.5}$$

When monthly values in a quarter are close to the monthly average over the quarter, which is typically true with seasonally adjusted figures, the approximation error introduced is negligible.

A.2 Variables in log differences

Using approximation (A.5) it is possible to write the quarter-on-quarter log difference $y_t = q_t - q_{t-3}$ as a function of the monthly log differences $x_t = m_t - m_{t-1}$ as

$$
\begin{aligned}
y_t &= \log(3) + \frac{1}{3}m_t + \frac{1}{3}m_{t-1} + \frac{1}{3}m_{t-2} - (\log(3) + \frac{1}{3}m_{t-3} + \frac{1}{3}m_{t-4} + \frac{1}{3}m_{t-5}), \\
&= \frac{1}{3}m_t - \frac{1}{3}m_{t-1} + \frac{1}{3}m_{t-1} + \frac{1}{3}m_{t-1} - \frac{2}{3}m_{t-2} + \frac{2}{3}m_{t-2} + \frac{1}{3}m_{t-2} \\
&- m_{t-3} + m_{t-3} - \frac{1}{3}m_{t-3} - \frac{2}{3}m_{t-4} + \frac{2}{3}m_{t-4} - \frac{1}{3}m_{t-4} - \frac{1}{3}m_{t-4} + \frac{1}{3}m_{t-4} - \frac{1}{3}m_{t-5} \\
&= \frac{1}{3}x_t + \frac{2}{3}x_{t-1} + x_{t-2} + \frac{2}{3}x_{t-3} + \frac{1}{3}x_{t-4}.
\end{aligned}
\tag{A.6}
$$

Equation (A.6) is also discussed by Mariano and Murasawa (2003) who have popularised its use in mixed-frequency dynamic factor models.

Appendix B

On the indeterminacy when trying to estimate monthly seasonal effects from rolling quarterly figures

The quarterly seasonal effects can be expressed as a linear function of the monthly seasonal effects as:

$$\begin{pmatrix} \text{Jan-Mar} \\ \text{Feb-Apr} \\ \text{Mar-May} \\ \text{Apr-Jun} \\ \text{May-Jul} \\ \text{Jun-Aug} \\ \text{Jul-Sep} \\ \text{Aug-Oct} \\ \text{Sep-Nov} \\ \text{Oct-Dec} \\ \text{Nov-Jan} \\ \text{Deb-Feb} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} \text{Jan} \\ \text{Feb} \\ \text{Mar} \\ \text{Apr} \\ \text{May} \\ \text{Jun} \\ \text{Jul} \\ \text{Aug} \\ \text{Sep} \\ \text{Oct} \\ \text{Nov} \\ \text{Dec} \end{pmatrix},$$

where Jan refers to the monthly seasonal effect for January and Jan-Mar refers to the quarterly seasonal effect for the period January to March. Two of the twelve rows of the linear system are linearly dependent of the other tens. For instance:

Row 11 = Row 1 + Row 4 + Row 7 + Row 10 - Row 2 - Row 5 - Row 8,

Row 12 = Row 1 + Row 4 + Row 7 + Row 10 - Row 3 - Row 6 - Row 9.

Consequently, the system is indeterminate and it is not possible to find unique monthly seasonal effects.

Appendix C

Linearisation with a SLC algorithm

The observation equation is linearised by taking the first order Taylor expansion of $\mathrm{Z}_t(.)$ at a guess $\tilde{\boldsymbol{\alpha}}_t$, which yields

$$\boldsymbol{y}_t = \mathrm{Z}_t(\tilde{\boldsymbol{\alpha}}_t) + \dot{\boldsymbol{Z}}_t.(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_t), \qquad \text{where} \quad \dot{\boldsymbol{Z}}_t = \left.\frac{\partial \mathrm{Z}_t(x)}{\partial x}\right|_{x=\tilde{\boldsymbol{\alpha}}_t}.$$

Since $\mathrm{Z}_t(\tilde{\boldsymbol{\alpha}}_t)$ and $\dot{\boldsymbol{Z}}_t.\tilde{\boldsymbol{\alpha}}_t$ are not random quantities, but instead are fixed prior to the estimation, a new observation vector $\tilde{\boldsymbol{y}}_t$ can be defined such that

$$\tilde{\boldsymbol{y}}_t = \dot{\boldsymbol{Z}}_t.\boldsymbol{\alpha}_t, \qquad \text{where} \quad \tilde{\boldsymbol{y}}_t = \boldsymbol{y}_t - \mathrm{Z}_t(\tilde{\boldsymbol{\alpha}}_t) + \dot{\boldsymbol{Z}}_t.\tilde{\boldsymbol{\alpha}}_t. \tag{C.1}$$

Replacing the observation equation with the approximated linear observation equation (C.1) gives the approximated linear model

$$\begin{aligned} \tilde{\boldsymbol{y}}_t &= \dot{\boldsymbol{Z}}_t.\boldsymbol{\alpha}_t, \\ \boldsymbol{\alpha}_{t+1} &= T\boldsymbol{\alpha}_t + R\boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t \sim \mathrm{N}(0, Q), \\ \boldsymbol{\alpha}_1 &\sim \mathrm{N}(\boldsymbol{a}_1, P_1), \end{aligned} \tag{C.2}$$

which can be estimated using the standard Kalman filter and smoother. It is now possible to set out the SLC algorithm following Proietti and Moauro (2006):

1: Generate model (C.2) at a guess $\tilde{\boldsymbol{\alpha}}_t$;

2: Run the Kalman filter and smoother on model (C.2), which yields the smoothed state vector $\hat{\boldsymbol{\alpha}}_t$;

3: Set $\tilde{\boldsymbol{\alpha}}_t$ to $\hat{\boldsymbol{\alpha}}_t$;

4: Iterate steps 1 to 3 until $\hat{\boldsymbol{\alpha}}_t = \tilde{\boldsymbol{\alpha}}_t$.

Once the algorithm has converged the following equation holds

$$\begin{aligned}\boldsymbol{y}_t &= \mathrm{Z}_t(\tilde{\boldsymbol{\alpha}}_t) + \dot{\boldsymbol{Z}}_t.(\hat{\boldsymbol{\alpha}}_t - \tilde{\boldsymbol{\alpha}}_t)\\ &= \mathrm{Z}_t(\hat{\boldsymbol{\alpha}}_t)\end{aligned}$$

where the approximation error is reduced to zero.

Appendix D

Full matrix representation of the state space models

D.1 Bivariate model for temporal disaggregation

The state space representation (3.11) is

$$
\begin{aligned}
y_t &= Z_t \alpha_t, \\
\alpha_{t+1} &= T\alpha_t + R\eta_t, \qquad \eta_t \sim \mathrm{N}(0, Q), \\
\alpha_1 &\sim \mathrm{N}(a_1, P_1).
\end{aligned}
$$

$0_{m\times n}$ is a matrix of zeros with m rows and n columns. The observation matrix is $Z_t = (Z_{1,t}, Z_{2,t})'$, where
$Z_{1,t} = (1/3, 1/3, 1/3, 0, 1, 0_{10\times 1}, 1/3, 1/3, 1/3, h^a_{1,t}, 1^{(2)}_{b,t}, 1^{(3)}_{b,t}, 1^{(1)}_{\epsilon,t}, 1^{(2)}_{\epsilon,t}, 1^{(3)}_{\epsilon,t}, 0_{14\times 1})$,
$Z_{2,t} = (0_{23\times 1}, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, h^a_{2,t})$. The variables $1^{(j)}_{b,t}$, $j = 2, 3$, are stagger dummy variables relating to the stagger biases, while $1^{(j)}_{\epsilon,t}$, $j = 1, 2, 3$, are stagger dummy variables relating to the observation errors.

The state vector is $\alpha_t = (\alpha'_{1,t}, \alpha'_{2,t})'$, where

$\alpha_{1,t} = (\mu_{1,t}, \mu_{1,t-1}, \mu_{1,t-2}, \nu_{1,t}, \gamma_{1,t}, ..., \gamma_{1,t-8}, e_{1,t}, e_{1,t-1}, e_{1,t-2}, \beta_1, b^{(2)}_{1,t}, b^{(3)}_{1,t}, \epsilon^{(1)}_{1,t}, \epsilon^{(2)}_{1,t}, \epsilon^{(3)}_{1,t})'$,

$\alpha_{2,t} = (\mu_{2,t}, \nu_{2,t}, \gamma_{2,1,t}, \gamma^*_{2,1,t}, \gamma_{2,2,t}, \gamma^*_{2,2,t}, ..., , \gamma_{2,5,t}, \gamma^*_{2,5,t}, \gamma_{2,6,t}, e_{2,t}, \beta_2)'$.

The transition matrix is $T = \text{diag}(T_{1,LLT}, T_{1,\gamma}, T_{1,e}, T_{1,\beta}, T_{1,b}, T_{1,\epsilon}, T_{2,LLT}, T_{2,\gamma}, T_{2,e}, T_{2\beta})$,

where $T_{1,LLT} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$; $T_{1,\gamma} = \begin{pmatrix} 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$;

$T_{1,e} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$; $T_{1,\beta} = 1$, $T_{1,b} = I_2$; $T_{1,\epsilon} = 0_{3\times3}$; $T_{2,LLT} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$; $T_{2,e} = 0$

$T_{2,\gamma} = \text{diag}(T_{2,1,\gamma}, T_{2,2,\gamma}, T_{2,3,\gamma}, T_{2,4,\gamma}, T_{2,5,\gamma}, -1)$; $T_{2,j,\gamma} = \begin{pmatrix} \cos(2\pi j/12) & \sin(2\pi j/12) \\ -\sin(2\pi j/12) & \cos(2\pi j/12) \end{pmatrix}$;

$T_{2,\beta} = 1$. $R = \text{diag}(R_{1,\mu}, R_{1,\nu}, R_{1,e}, R_{1,b}, R_{1,\epsilon}, R_{2,\mu}, R_{2,\nu}, R_{2,\gamma}, R_{2,e})$, where $R_{1,\mu} = R_{1,e} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}'$; $R_{1,\nu} = 1$; $R_{1,b} = I_2$;$R_{1,\epsilon} = I_3$ $R_{2,\mu} = R_{2,\nu} = R_{2,e} = 1$; $R_{2,\gamma} = I_{11}$.

Q is a 22×22 matrix with diagonal $(\sigma^2_{1,\xi}, \sigma^2_{1,\zeta}, \sigma^2_{1,e}, \sigma^2_{1,\kappa2}, \sigma^2_{1,\kappa3}, \sigma^2_{1,\epsilon1}, \sigma^2_{1,\epsilon2}, \sigma^2_{1,\epsilon3}, \sigma^2_{2,\xi}, \sigma^2_{2,\zeta}, \sigma^2_{2,\gamma}, ..., \sigma^2_{2,\gamma}, \sigma^2_{2,\gamma}/2, \sigma^2_{2,e})$, and $Q_{[1,5]} = \rho_\xi \sigma_{1,\xi}\sigma_{2,\xi}$, $Q_{[2,6]} = \rho_\zeta \sigma_{1,\zeta}\sigma_{2,\zeta}$, $Q_{[3,7]} = \rho_e \sigma_{1,e}\sigma_{2,e}$.

For estimation, Q is expressed using the Cholesky factorisation as $Q = LL'$ and minimise the negative log likelihood function with respect to the parameters in the lower triangular matrix L. Thus the estimated variance-covariance matrix Q is positive semi-definite.

D.2 Multivariate model for nowcasting

The state space representation (3.11) is

$$\begin{aligned} y_t &= Z_t \alpha_t, \\ \alpha_{t+1} &= T\alpha_t + R\eta_t, \qquad \eta_t \sim \mathrm{N}(0, Q), \\ \alpha_1 &\sim \mathrm{N}(a_1, P_1). \end{aligned}$$

$0_{m \times n}$ as a matrix of zeros with m rows and n columns. The observation matrix is $Z_t = (Z_{0,t}, Z_n, Z_\chi, Z_c)$, where

$Z_{0,t} = (Z_{1,t}, Z_{1,t}, Z_{1,t}, Z_{1,t}, Z_{1,t}, Z_{1,t}, Z_{1,t}, Z_{1,t}, Z_{1,t}, Z_{1,t}, Z_{1,t}, Z_{2,t})'$;

$Z_{1,t} = (1/3, 1/3, 1/3, 0, 1, 0_{8\times1}, 0, 1/3, 1/3, 1/3, h^a_{1,t}, 1^1_{b,t}, 1^2_{b,t}, 1^3_{b,t}, 1^1_{\chi,t}, 1^2_{\chi,t}, 1^3_{\chi,t}, 0_{14\times1})$,

$Z_{2,t} = (0_{24\times1}, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, h^a_{2,t})$. The variables $1^j_{b,t}$, $j = 1, 2, 3$, are stagger dummy variables relating to the stagger biases, while $1^j_{\chi,t}$, $j = 1, 2, 3$, are stagger dummy variables relating to the persistent measurement errors linked to the eleventh release (which therefore affect all releases).

The state vector is $\alpha_t = (\alpha'_{1,t}, \alpha'_{2,t})'$, where

$\alpha_{1,t} = (\mu_{1,t}, \mu_{1,t-1}, \mu_{1,t-2}, \nu_{1,t}, \gamma_{1,t}, ..., \gamma_{1,t-8}, e_{1,t}, e_{1,t-1}, e_{1,t-2}, \beta_1, b^1_{1,t}, b^2_{1,t}, b^3_{1,t}, \chi^1_{1,t}, \chi^2_{1,t}, \chi^3_{1,t})'$,

$\alpha_{2,t} = (\mu_{2,t}, \nu_{2,t}, \gamma_{2,1,t}, \gamma^*_{2,1,t}, \gamma_{2,2,t}, \gamma^*_{2,2,t}, ...,, \gamma_{2,5,t}, \gamma^*_{2,5,t}, \gamma_{2,6,t}, e_{2,t}, \beta_2)'$.

$$Z_n = \begin{pmatrix} 1&1&1&1&1&1&1&1&1&1\\ 0&1&1&1&1&1&1&1&1&1\\ 0&0&1&1&1&1&1&1&1&1\\ 0&0&0&1&1&1&1&1&1&1\\ 0&0&0&0&1&1&1&1&1&1\\ 0&0&0&0&0&1&1&1&1&1\\ 0&0&0&0&0&0&1&1&1&1\\ 0&0&0&0&0&0&0&1&1&1\\ 0&0&0&0&0&0&0&0&1&1\\ 0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0 \end{pmatrix};\ Z_\epsilon = \begin{pmatrix} 0&0&0&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&0&0&1\\ 0&0&0&0&0&0&0&0&0&0&0 \end{pmatrix};\ Z_c = \begin{pmatrix} 1&0&0&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0&0&0&0&0\\ 0&0&0&0&0&1&0&0&0&0&0\\ 0&0&0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0 \end{pmatrix}.$$

The transition matrix is $T = \text{diag}(T_{1,LLT}, T_{1,\gamma}, T_{1,e}, T_{1,\beta}, T_{1,b}, T_{1,\epsilon}, T_{2,LLT}, T_{2,\gamma}, T_{2,e}, T_{2\beta})$, where $T_{1,LLT} = \begin{pmatrix} 1&0&0&1\\ 1&0&0&0\\ 0&1&0&0\\ 0&0&0&1 \end{pmatrix}$; $T_{1,\gamma} = \begin{pmatrix} 0&0&-1&0&0&-1&0&0&-1\\ 1&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&0&0\\ 0&0&0&0&1&0&0&0&0\\ 0&0&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&1&0 \end{pmatrix}$;

$T_{1,e} = \begin{pmatrix} 0&0&0\\ 1&0&0\\ 0&1&0 \end{pmatrix}$; $T_{1,\beta} = 1$, $T_{1,b} = I_2$; $T_{1,\epsilon} = 0_{3\times 3}$; $T_{2,LLT} = \begin{pmatrix} 1&1\\ 0&1 \end{pmatrix}$; $T_{2,e} = 0$

$T_{2,\gamma} = \text{diag}(T_{2,1,\gamma}, T_{2,2,\gamma}, T_{2,3,\gamma}, T_{2,4,\gamma}, T_{2,5,\gamma}, -1)$; $T_{2,j,\gamma} = \begin{pmatrix} \cos(2\pi j/12) & \sin(2\pi j/12)\\ -\sin(2\pi j/12) & \cos(2\pi j/12) \end{pmatrix}$; $T_{2,\beta} = 1$. $R = \text{diag}(R_{1,\mu}, R_{1,\nu}, R_{1,e}, R_{1,b}, R_{1,\epsilon}, R_{2,\mu}, R_{2,\nu}, R_{2,\gamma}, R_{2,e})$, $T_\epsilon = 0_{10\times 10}$ $T_\chi = 0_{10\times 10}$, $T_c = 1_{10\times 10}$, where $R_{1,\mu} = R_{1,e} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}'$; $R_{1,\nu} = 1$; $R_{1,b} = I_2$;$R_{1,\epsilon} = I_3$ $R_{2,\mu} = R_{2,\nu} = R_{2,e} = 1$; $R_{2,\gamma} = I_{11}$.

Q is a 23×23 matrix with diagonal $(\sigma^2_{1,\xi}, \sigma^2_{1,\zeta}, \sigma^2_{1,e}, \sigma^2_{1,\kappa 2}, \sigma^2_{1,\kappa 3}, \sigma^2_{1,\epsilon 1}, \sigma^2_{1,\epsilon 2}, \sigma^2_{1,\epsilon 3}, \sigma^2_{2,\xi}, \sigma^2_{2,\zeta}, \sigma^2_{2,\omega}, ..., \sigma^2_{2,\omega}, \sigma^2_{2,\omega}/2, \sigma^2_{2,e})$, and $Q_{[1,5]} = \rho_\xi \sigma_{1,\xi} \sigma_{2,\xi}$, $Q_{[2,6]} = \rho_\zeta \sigma_{1,\zeta} \sigma_{2,\zeta}$, $Q_{[3,7]} = \rho_e \sigma_{1,e} \sigma_{2,e}$.

For estimation, Q is expressed using the Cholesky factorisation as $Q = LL'$ and minimise the negative log likelihood function with respect to the parameters in the lower triangular matrix L. Thus, the estimated variance-covariance matrix Q is positive semi-definite.

Appendix E

Elements of the score and information matrix

This appendix provides the diagonal of the expected information matrix

$$\mathrm{E}\left[\left(\frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial a_{i,t}}\right)' \frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial a_{i,t}} |Y_{t-1}\right],$$

defined as $\mathcal{I}^{\mu}_{i,t}$, $\mathcal{I}^{\sigma}_{i,t}$ and $\mathcal{I}^{\alpha}_{i,t}$, as well as the score elements necessary for their computation. The first derivative of the log density (5.28) with respect to location parameters is

$$\Delta^{\mu}_{i,t} = \frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial \mu_{i,t}} = \frac{\nu_{i,1}+1}{1+\frac{1}{\nu_{i,1}}\left(\frac{y_{i,t}-\mu_{i,t}}{2\alpha_{i,t}\sigma_{i,t}K(\nu_{i,1})}\right)^2} . \frac{y_{i,t}-\mu_{i,t}}{\nu_{i,1}(2\alpha_{i,t}\sigma_{i,t}K(\nu_{i,1}))^2} \; 1(y_{i,t} \leq \mu_{i,t})$$
$$+ \frac{\nu_{i,2}+1}{1+\frac{1}{\nu_{i,2}}\left(\frac{y_{i,t}-\mu_{i,t}}{2(1-\alpha_{i,t})\sigma_{i,t}K(\nu_{i,2})}\right)^2} . \frac{y_{i,t}-\mu_{i,t}}{\nu_{i,2}(2(1-\alpha_{i,t})\sigma_{i,t}K(\nu_{i,2}))^2} \; 1(y_{i,t} > \mu_{i,t}). \tag{E.1}$$

The elements of the information matrix corresponding to location parameters are given by

$$\mathcal{I}^{\mu}_{i,t} = \mathrm{E}\big[\Delta^{\mu}_{i,t}\Delta^{\mu}_{i,t}|Y_{t-1}\big] = \frac{1}{\sigma^2_{i,t}}\Big[\frac{\nu_{i,1}+1}{\alpha_{i,t}(\nu_{i,1}+3)K^2(\nu_{i,1})} + \frac{\nu_{i,2}+1}{(1-\alpha_{i,t})(\nu_{i,2}+3)K^2(\nu_{2,i})}\Big]. \tag{E.2}$$

The first derivative of the log density (5.28) with respect to scale parameters is

$$\begin{aligned}\Delta^{\sigma}_{i,t} = \frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial \sigma_{i,t}} &= \big[\frac{(\nu_{i,1}+1)}{1+(\frac{y_{i,t}-\mu_{i,t}}{2\alpha_{i,t}K(\nu_{i,1})\sigma_{i,t}\sqrt{\nu_{i,1}}})^2} \\ &\times (\frac{y_{i,t}-\mu_{i,t}}{2\alpha_{i,t}\sigma_{i,t}K(\nu_{i,1})\sqrt{\nu_{i,1}}})^2 - 1\big]/\sigma_{i,t}\ 1(y_{i,t} < \mu_{i,t}), \\ &+ \big[(\nu_{i,2}+1)\frac{1}{1+(\frac{y_{i,t}-\mu_{i,t}}{2(1-\alpha_{i,t})K(\nu_{i,2})\sigma_{i,t}\sqrt{\nu_{i,2}}})^2} \\ &\times \frac{y_{i,t}-\mu_{i,t}}{2(1-\alpha_{i,t})\sigma_{i,t}K(\nu_{i,2})\sqrt{\nu_{i,2}}})^2 - 1\big]/\sigma_{i,t}\ 1(y_{i,t} > \mu_{i,t}).\end{aligned} \tag{E.3}$$

The elements of the information matrix corresponding to scale parameters are given by

$$\mathcal{I}^{\sigma}_{i,t} = \mathrm{E}\big[\Delta^{\sigma}_{i,t}\Delta^{\sigma}_{i,t}|Y_{t-1}\big] = \frac{2}{\sigma^2_{i,t}}\Big[\frac{\alpha_{i,t}\nu_{i,1}}{\nu_{i,1}+3} + \frac{(1-\alpha_{i,t})\nu_{i,2}}{\nu_{i,2}+3}\Big]. \tag{E.4}$$

The first derivative of the log density (5.28) with respect to shape parameters is

$$\begin{aligned}\Delta^{\alpha}_{i,t} = \frac{\partial \mathrm{log} f_i(y_{i,t}|Y_{t-1})}{\partial \alpha_{i,t}} &= \frac{\nu_{i,1}+1}{\nu_{i,1}}\frac{1}{1+(\frac{y_{i,t}-\mu_{i,t}}{2\alpha_{i,t}K(\nu_{i,1})\sigma_{i,t}\sqrt{\nu_{i,1}}})^2} \\ &\times ((\frac{y_{i,t}-\mu_{i,t}}{2\sigma_{i,t}K(\nu_{i,1})\sqrt{\nu_{i,1}}})^2\frac{1}{\alpha^3_{i,t}}\ 1(y_{i,t} < \mu_{i,t}), \\ &+ \frac{\nu_{i,2}+1}{\nu_{i,2}}\frac{1}{1+(\frac{y_{i,t}-\mu_{i,t}}{2(1-\alpha_{i,t})K(\nu_{i,2})\sigma_{i,t}\sqrt{\nu_{i,2}}})^2} \\ &\times ((\frac{y_{i,t}-\mu_{i,t}}{2\sigma_{i,t}K(\nu_{i,2})\sqrt{\nu_{i,2}}})^2\frac{1}{(1-\alpha_{i,t})^3}\ 1(y_{i,t} > \mu_{i,t}).\end{aligned} \tag{E.5}$$

The elements of the information matrix corresponding to shape parameters are given by

$$\mathcal{I}^{\alpha}_{i,t} = \mathrm{E}\big[\Delta^{\alpha}_{i,t}\Delta^{\alpha}_{i,t}|Y_{t-1}\big] = 3\Big[\frac{\nu_{i,1}+1}{\alpha_{i,t}(\nu_{i,1}+3)} + \frac{\nu_{i,2}+1}{(1-\alpha_{i,t})(\nu_{i,2}+3)}\Big]. \tag{E.6}$$

The formulae for the information matrix can be found in Zhu and Galbraith (2010).