A Data–Cleaning Augmented Kalman Filter for Robust Estimation of State Space Models*

Martyna Marczak[a,*], Tommaso Proietti[b,c,1], Stefano Grassi[d]

[a] *University of Hohenheim, Germany*
[b] *Università di Roma Tor Vergata, Italy*
[c] *CREATES, Denmark*
[d] *University of Kent, United Kingdom*

Abstract

This article presents a robust augmented Kalman filter that extends the data-cleaning filter [Masreliez, C.J., and Martin, R.D., *IEEE Trans. Autom. Control, AC-22*, 361–371, 1977] to the general state space model featuring nonstationary and regression effects. The robust filter shrinks the observations towards their one-step-ahead prediction based on the past, by bounding the effect of the information carried by a new observation according to an influence function. When maximum likelihood estimation is carried out on the replacement data, an M-type estimator is obtained. The performance of the robust AKF is investigated in two applications using as a modeling framework the basic structural time series model, a popular unobserved components model in the analysis of seasonal time series. First, a Monte Carlo experiment is conducted in order to evaluate the comparative accuracy of the proposed method for estimating the variance parameters. Second, the method is applied in a forecasting context to a large set of European trade statistics series.

Keywords: Robust Filtering, Augmented Kalman filter, Structural time series model, Additive outlier, Innovation outlier

*Supplementary material can be found in the online appendix to this article.

*Corresponding author at: University of Hohenheim, Department of Economics, Schloss Hohenheim 1C, D-70593 Stuttgart, Germany

Email addresses: marczak@uni-hohenheim.de (Martyna Marczak), tommaso.proietti@uniroma2.it (Tommaso Proietti), S.Grassi@kent.ac.uk (Stefano Grassi)

[1]Tommaso Proietti acknowledges support from CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), funded by the Danish National Research Foundation.

1. Introduction

State space models and the Kalman filter offer a powerful tool for statistical analysis of time series. Any linear Markovian time series model can be put into a state space form and then the Kalman filter can be applied to estimate the model parameters by maximum likelihood via the prediction error decomposition. However, if the considered series is contaminated by outliers, the estimated parameters might be strongly biased. Masreliez and Martin (1977) proposed a robustification of the Kalman filter to remove the outlier effects. This approach relies on scaling residuals by an influence function, a continuous and bounded function, so that the estimator resulting from the application of the robust Kalman filter belongs to the class of M-estimators. Martin (1979) and Martin and Thomson (1982) are early works examining the robust Kalman filter in the case of ARMA models. A comprehensive account is provided in chapter 8 of Maronna et al. (2006). Recent extensions of the original robust Kalman filter can be found in works by, e.g., Liu et al. (2004), Gandhi and Mili (2010) and Ruckdeschel et al. (2014), which cover both stationary and nonstationary settings.

In this paper we propose another extension of the robust Kalman filter which is suited for nonstationary series and/or models capturing regressor effects. In particular, we build on the augmented Kalman filter (AKF) proposed by de Jong (1991). To the best of our knowledge, this is the first paper that combines the AKF with a robustification procedure.

Our approach is based on the heuristic argument of shrinking a suspect observation towards its one-step-ahead prediction so as to achieve robustness. A theoretically consistent and empirically viable approach to robustness in time series analysis has been recently proposed by Harvey (2013). The approach, however, deals with unobserved components whose dynamics is driven by the conditional score of the observation density and, unlike our proposed method, cannot handle models with multiple source of errors.

Since the presented approach—the robust AKF—might be of relevance for economic time series many of which are nonstationary and affected by outlying observations, we consider the class of structural time series models, i.e. models formulated in terms of unobserved components, like trend, cycle, seasonal or irregular components, inherent in many economic time series (Harvey, 1989). To account for the fact that treatment of outliers usually accompanies the removal of the seasonal component from the data, as the reference model we use the basic structural model (BSM) for univariate series (Harvey and Todd, 1983), often applied for the purpose of seasonal adjustment. The BSM is a simple yet flexible model providing a satisfactory fit to a wide range of seasonal time series.

Using the BSM as the modeling framework, we investigate the performance of the

robust AKF in a Monte Carlo simulation exercise and in an empirical application. In the Monte Carlo experiment, the simulated series are affected by additive or innovation outliers. Our results complement and extend Bianco et al. (2001), in that we assess the role of specific outlier types on estimating the correct size of a particular variance component in a nonstationary dynamic regression framework. In the empirical application, we conduct a pseudo real–time forecasting experiment on a large dataset of 540 monthly European trade statistics series, that are contaminated with additive outliers. The aim of the study is to investigate whether the robust AKF is capable of cleaning the data effectively and thus reducing the forecast uncertainty.

The remainder of the article is organized as follows. In Section 2, we set out the reference state space form and we present the AKF. The robust AKF is exposed in Section 3. Section 4 presents the modeling framework used in the application part of the study. In Section 5, we evaluate the robust AKF by means of a Monte Carlo experiment whereas in Section 6 we discuss the results of the forecasting exercise. Section 7 concludes.

2. State Space Models

The augmented Kalman filter, see Rosenberg (1973) and de Jong (1991), is an essential algorithm for likelihood inferences on the parameters of a state space model and for linear prediction. Given the parameter values, it evaluates the likelihood via the prediction error decomposition, and once the parameters are estimated as the maximisers of the likelihood function, it enables the out-of-sample prediction of the series and the estimation of the states in real time.

2.1. Specification

Consider a multivariate time series $\mathbf{y}_t$ with N elements, observed at times $t = 1, \ldots, n$. The state space model for $\mathbf{y}_t$ is formulated as follows:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{Z}_t\boldsymbol{\alpha}_t + \mathbf{X}_t\boldsymbol{\beta} + \mathbf{G}_t\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \text{IIDN}(\mathbf{0}, \sigma^2\mathbf{I}), \\ \boldsymbol{\alpha}_{t+1} &= \mathbf{T}_t\boldsymbol{\alpha}_t + \mathbf{W}_t\boldsymbol{\beta} + \mathbf{H}_t\boldsymbol{\varepsilon}_t, \quad t = 1, ..., n, \end{aligned} \tag{1}$$

where $\boldsymbol{\alpha}_t$ is an $(m \times 1)$ vector of states, $\mathbf{Z}_t$ is an $N \times m$ matrix, $\mathbf{X}_t$ denotes an $(N \times k)$ matrix, $\boldsymbol{\beta}$ is a $(k \times 1)$ vector, $\mathbf{G}_t$ is $(N \times r)$, and $\boldsymbol{\varepsilon}_t$ is an $(r \times 1)$ vector of independently and identically distributed (IID) Gaussian random variable with unit covariance matrix. The second equation, referred to as the *transition equation*, is a first order Markovian model for the evolution of the state vector; $\mathbf{T}_t$ is a $(m \times m)$ matrix and $\mathbf{H}_t$ is a $(m \times r)$ matrix, and $\mathbf{W}_t$ is $(m \times k)$. The system matrices $\mathbf{Z}_t, \mathbf{G}_t, \mathbf{T}_t, \mathbf{H}_t$ are non-stochastic, can be time varying, and in general contain unknown parameters denoted by $\boldsymbol{\theta}$, referred to as

hyperparameters, that have to be estimated along with the fixed effects in $\boldsymbol{\beta}$; $\mathbf{X}_t$ and $\mathbf{W}_t$ denote contain the exogenous measurements.

The initial state vector is specified as follows:

$$\boldsymbol{\alpha}_1 = \tilde{\boldsymbol{\alpha}}^*_{1|0} + \mathbf{W}_0\boldsymbol{\beta} + \mathbf{H}_0\boldsymbol{\varepsilon}_0, \tag{2}$$

where $\tilde{\boldsymbol{\alpha}}^*_{1|0}$, $\mathbf{W}_0$, and $\mathbf{H}_0$ are known quantities.

The leading case of interest is when $\boldsymbol{\beta}$ is partitioned as

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\alpha}_0^\dagger \\ \boldsymbol{\beta}_x \\ \boldsymbol{\beta}_w \end{bmatrix}, \quad \begin{aligned} \mathbf{X}_t &= \left[\mathbf{0}, \mathbf{X}_t^\dagger, \mathbf{0}\right], \\ \mathbf{W}_t &= \left[\mathbf{0}, \mathbf{0}, \mathbf{W}_t^\dagger\right], \\ \mathbf{W}_0 &= \left[\mathbf{T}^\dagger, \mathbf{0}, \mathbf{W}_0^\dagger\right], \end{aligned}$$

where $\boldsymbol{\alpha}_0^\dagger$ are a subset of initial states corresponding to nonstationary elements of $\boldsymbol{\alpha}_t$, $\mathbf{X}_t^\dagger$ is an $(N \times k_x)$ matrix of explanatory variables affecting the response variable, $\mathbf{W}_t^\dagger$ is an $(m \times k_w)$ matrix of explanatory variables affecting $\boldsymbol{\alpha}_{t+1}$, and $\mathbf{T}^\dagger$ is a matrix relating $\boldsymbol{\alpha}_1$ to $\boldsymbol{\alpha}_0^\dagger$.

The vector $\boldsymbol{\beta}$ can be considered as fixed (and unknown), or as a random vector with a diffuse distribution, $\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \kappa\mathbf{I})$, where κ is an arbitrarily large number, tending to ∞. A diffuse distribution for $\boldsymbol{\beta}$ is used to quantify our uncertainty about the effects of the exogenous variables and the initial conditions.

2.2. The Augmented Kalman Filter

Consider the state space model (1), with initial conditions stated in (2). Setting $\mathbf{A}_{1|0} = -\mathbf{W}_0$, $\mathbf{P}^*_{1|0} = \mathbf{H}_0\mathbf{H}_0'$, the AKF is, for $t = 1, \ldots, n$:

$$\begin{aligned} &\boldsymbol{\nu}^*_t = \mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}^*_{t|t-1}, && \mathbf{V}_t = \mathbf{X}_t - \mathbf{Z}_t\mathbf{A}_{t|t-1}, \\ &\mathbf{F}^*_t = \mathbf{Z}_t\mathbf{P}^*_{t|t-1}\mathbf{Z}_t' + \mathbf{G}_t\mathbf{G}_t', && \mathbf{K}^*_t = (\mathbf{T}_t\mathbf{P}^*_{t|t-1}\mathbf{Z}_t' + \mathbf{H}_t\mathbf{G}_t')\mathbf{F}_t^{*-1}, \\ &\tilde{\boldsymbol{\alpha}}^*_{t+1|t} = \mathbf{T}_t\tilde{\boldsymbol{\alpha}}^*_{t|t-1} + \mathbf{K}^*_t\boldsymbol{\nu}^*_t, && \mathbf{A}_{t+1|t} = \mathbf{T}_t\mathbf{A}_{t|t-1} - \mathbf{W}_t + \mathbf{K}^*_t\mathbf{V}_t, \\ &\mathbf{P}^*_{t+1|t} = \mathbf{T}_t\mathbf{P}^*_{t|t-1}\mathbf{T}_t' + \mathbf{H}_t\mathbf{H}_t' - \mathbf{K}^*_t\mathbf{F}^*_t\mathbf{K}^{*\prime}_t. \end{aligned} \tag{3}$$

The starred quantities correspond to the usual Kalman filter applied to $\mathbf{y}_t$ with $\boldsymbol{\beta} = \mathbf{0}$, when the state vector is initialized by $\tilde{\boldsymbol{\alpha}}^*_{1|0}$ and no explanatory variables are considered. The vector $\boldsymbol{\nu}^*_t$ represents the innovations in such a case, $\boldsymbol{\nu}^*_t = \mathbf{y}_t - \mathrm{E}(\mathbf{y}_t|\boldsymbol{\mathcal{Y}}_{t-1}, \boldsymbol{\beta} = \mathbf{0})$, whereas $\mathbf{F}^*_t$ denotes their covariance matrix, $\mathbf{F}^*_t = \sigma^{-2}\mathrm{Var}(\mathbf{y}_t|\boldsymbol{\mathcal{Y}}_{t-1}, \boldsymbol{\beta} = \mathbf{0})$. Here, $\boldsymbol{\mathcal{Y}}_{t-1} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{t-1}\}$ is the past history of $\mathbf{y}_t$. The Kalman gain matrix has the following interpretation: $\mathbf{K}^*_t = \mathrm{Cov}(\boldsymbol{\alpha}_t, \mathbf{y}_t|\boldsymbol{\mathcal{Y}}_{t-1}, \boldsymbol{\beta} = \mathbf{0})[\mathrm{Var}(\mathbf{y}_t|\boldsymbol{\mathcal{Y}}_{t-1}, \boldsymbol{\beta} = \mathbf{0})]^{-1}$. The

vector $\tilde{\boldsymbol{\alpha}}^*_{t+1|t} = \mathrm{E}(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\mathcal{Y}}_t, \boldsymbol{\beta} = \mathbf{0})$ is the one–step–ahead prediction of the state vector given the information in period t and conditional on $\boldsymbol{\beta} = \mathbf{0}$. The corresponding conditional covariance matrix is $\mathbf{P}^*_{t+1|t}$. Notice that the quantities $\mathbf{F}^*_t, \mathbf{K}^*_t$ and $\mathbf{P}_{t+1|t}$ do not depend on the observation $\mathbf{y}_t$ (and neither upon $\mathbf{X}_t, \mathbf{W}_t$).

The matrix recursions for $\mathbf{V}_t$ and $\mathbf{A}_{t+1|t}$ are run in parallel to the above conditional Kalman filter and amount to running the Kalman filter to each of the columns of $\mathbf{X}_t$, so that, for instance, the innovations are $\mathbf{y}_t - \mathrm{E}(\mathbf{y}_t|\boldsymbol{\mathcal{Y}}_{t-1}, \boldsymbol{\mathcal{X}}_t, \boldsymbol{\mathcal{W}}_{t-1}) = \boldsymbol{\nu}^*_t - \mathbf{V}_t\boldsymbol{\beta}$ and the one step ahead predictions of the state variables is written as $\mathrm{E}(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\mathcal{Y}}_t, \boldsymbol{\mathcal{X}}_t, \boldsymbol{\mathcal{W}}_t) = \tilde{\boldsymbol{\alpha}}^*_{t+1|t} - \mathbf{A}_{t+1|t}\boldsymbol{\beta}$. Here, $\boldsymbol{\mathcal{X}}_t = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_t\}, \boldsymbol{\mathcal{W}}_t = \{\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_t\}$.

Given the availability of a sample of $t > k$ observations, $\{\boldsymbol{\mathcal{Y}}_t, \boldsymbol{\mathcal{X}}_t, \boldsymbol{\mathcal{W}}_t\}$, the vector $\boldsymbol{\beta}$ can be estimated by generalized least squares, as the minimizer of the criterion function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{t} (\boldsymbol{\nu}^*_i - \mathbf{V}_i\boldsymbol{\beta})' \mathbf{F}^{*-1}_i (\boldsymbol{\nu}^*_i - \mathbf{V}_i\boldsymbol{\beta}).$$

Defining

$$\mathbf{s}_t = \sum_{i=1}^{t} \mathbf{V}'_i \mathbf{F}^{*-1}_i \boldsymbol{\nu}^*_i, \qquad \mathbf{S}_t = \sum_{i=1}^{t} \mathbf{V}'_i \mathbf{F}^{*-1}_i \mathbf{V}_i,$$

for $t \geq k$ we obtain

$$\tilde{\boldsymbol{\beta}}_t = \mathbf{S}^{-1}_t \mathbf{s}_t.$$

The variance-covariance matrix of the regression coefficients (scaled by σ^{-2}) is $\mathbf{B}_t = \mathbf{S}^{-1}_t$. Also, the estimator of σ^2 based on t observations is

$$\hat{\sigma}^2_t = \frac{1}{tN} S(\tilde{\boldsymbol{\beta}}_t) = \frac{1}{tN} \left[\sum_{i=1}^{t} \boldsymbol{\nu}^{*\prime}_i \mathbf{F}^{*-1}_i \boldsymbol{\nu}^*_i - \mathbf{s}'_t \mathbf{S}^{-1}_t \mathbf{s}_t \right].$$

It can be shown, see de Jong (1991), that in the diffuse case, $\tilde{\boldsymbol{\beta}}_t$ is the mean of the posterior distribution of $\boldsymbol{\beta}$ given a sample of t observations, whereas $\sigma^2 \mathbf{S}^{-1}_t$ is the posterior covariance matrix.

Replacing $\boldsymbol{\beta}$ by its estimate, we obtain the innovations $\boldsymbol{\nu}_t = \mathbf{y}_t - \mathrm{E}(\mathbf{y}_t|\boldsymbol{\mathcal{Y}}_{t-1}, \boldsymbol{\mathcal{X}}_t, \boldsymbol{\mathcal{W}}_{t-1})$, the one–step–ahead prediction of the state vector, $\tilde{\boldsymbol{\alpha}}_{t|t-1} = \mathrm{E}(\boldsymbol{\alpha}_t|\boldsymbol{\mathcal{Y}}_{t-1}, \boldsymbol{\mathcal{X}}_{t-1}, \boldsymbol{\mathcal{W}}_{t-1})$, and the corresponding estimation error covariance matrices, as follows:

$$\begin{array}{lcllcl} \boldsymbol{\nu}_t & = & \boldsymbol{\nu}^*_t - \mathbf{V}_t\tilde{\boldsymbol{\beta}}_{t-1}, & \mathbf{F}_t & = & \mathbf{F}^*_t + \mathbf{V}_t\mathbf{B}_{t-1}\mathbf{V}'_t, \\ \tilde{\boldsymbol{\alpha}}_{t|t-1} & = & \tilde{\boldsymbol{\alpha}}^*_{t|t-1} - \mathbf{A}_{t|t-1}\tilde{\boldsymbol{\beta}}_{t-1}, & \mathbf{P}_{t|t-1} & = & \mathbf{P}^*_{t|t-1} + \mathbf{A}_{t|t-1}\mathbf{B}_{t-1}\mathbf{A}'_{t|t-1}. \end{array} \tag{4}$$

2.3. Estimation of model hyperparameters

The system matrices in (1) $(\mathbf{Z}_t, \mathbf{G}_t, \mathbf{T}_t, \mathbf{H}_t)$ are functionally related to a set of hyperparameters $\boldsymbol{\theta}$. Their estimation is carried out by maximising the profile diffuse likelihood

$$L(\boldsymbol{\theta}) = -\frac{1}{2}\left[N(n-k)(\ln\hat{\sigma}^2 + 1) + \sum_{t=1}^{n} \ln|\mathbf{F}_t^*| + \ln|\mathbf{S}_n|\right], \tag{5}$$

where $\hat{\sigma}^2$ is the estimator of the scale parameter σ^2:

$$\hat{\sigma}^2 = \frac{1}{N(n-k)}\left[\sum_{t=1}^{n} \boldsymbol{\nu}_t^{*\prime}\mathbf{F}_t^{*-1}\boldsymbol{\nu}_t^* - \mathbf{s}_n'\mathbf{S}_n^{-1}\mathbf{s}_n\right]. \tag{6}$$

The notion of a diffuse likelihood is close to that of a marginal likelihood, being based on reduced rank linear transformation of the series that eliminates dependence on $\boldsymbol{\beta}$; see de Jong (1991) and Francke et al. (2010).

2.4. Real–time estimates and predictions

As $\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{V}_t'\mathbf{F}_t^*\mathbf{V}_t$, by the Sherman–Woodbury–Morrison matrix inversion lemma (Henderson and Searle, 1981),

$$\mathbf{S}_t^{-1} = \mathbf{S}_{t-1}^{-1} - \mathbf{S}_{t-1}^{-1}\mathbf{V}_t'\left(\mathbf{F}_t^* + \mathbf{V}_t\mathbf{S}_{t-1}^{-1}\mathbf{V}_t'\right)^{-1}\mathbf{V}_t\mathbf{S}_{t-1}^{-1}.$$

The updated estimate of the vector $\boldsymbol{\beta}$ is $\tilde{\boldsymbol{\beta}}_{t|t} = \mathbf{S}_t^{-1}\mathbf{s}_t$, which, in view of $\mathbf{s}_t = \mathbf{s}_{t-1} + \mathbf{V}_t'\mathbf{F}_t^*\boldsymbol{\nu}_t^*$ and the above matrix inverse, and recalling (4), can be written, after some algebra:

$$\tilde{\boldsymbol{\beta}}_t = \tilde{\boldsymbol{\beta}}_{t-1} + \mathbf{B}_{t-1}\mathbf{V}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t. \tag{7}$$

The updated covariance matrix is $\sigma^2\mathbf{B}_t$, where

$$\mathbf{B}_t = \mathbf{B}_{t-1} - \mathbf{B}_{t-1}\mathbf{V}_t'\mathbf{F}_t^{-1}\mathbf{V}_t\mathbf{B}_{t-1}. \tag{8}$$

The real-time estimates of the state vector, $\tilde{\boldsymbol{\alpha}}_{t|t} = \hat{E}(\boldsymbol{\alpha}_t|\boldsymbol{\mathcal{Y}}_t, \boldsymbol{\mathcal{X}}_t, \boldsymbol{\mathcal{W}}_t)$, and their covariance matrix $\mathrm{Var}(\boldsymbol{\alpha}_t|\boldsymbol{\mathcal{Y}}_t, \boldsymbol{\mathcal{X}}_t, \boldsymbol{\mathcal{W}}_t) = \sigma^2\mathbf{P}_{t|t}$ are obtained as:

$$\begin{aligned}
\tilde{\boldsymbol{\alpha}}_{t|t} &= \tilde{\boldsymbol{\alpha}}_{t|t-1}^* - \mathbf{A}_{t|t-1}\tilde{\boldsymbol{\beta}}_t + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1}(\boldsymbol{\nu}_t^* - \mathbf{V}_t\tilde{\boldsymbol{\beta}}_t), \\
\mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1}^* - \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1}\mathbf{Z}_t\mathbf{P}_{t|t-1}^* + (\mathbf{A}_{t|t-1} + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1})\mathbf{B}_t(\mathbf{A}_{t|t-1} + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1})'.
\end{aligned} \tag{9}$$

We shall refer to $\boldsymbol{\nu}_t^* - \mathbf{V}_t\tilde{\boldsymbol{\beta}}_t$ as the GLS residual. It is related to the innovations

according to

$$\boldsymbol{\nu}_t^* - \mathbf{V}_t\tilde{\boldsymbol{\beta}}_t = \mathbf{F}_t^*\mathbf{F}_t^{-1}\boldsymbol{\nu}_t,$$

so that we rewrite:

$$\tilde{\boldsymbol{\alpha}}_{t|t} = \tilde{\boldsymbol{\alpha}}_{t|t-1}^* - \mathbf{A}_{t|t-1}\tilde{\boldsymbol{\beta}}_t + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t. \tag{10}$$

Furthermore, denoting the real-time estimates of the disturbance vector by $\tilde{\boldsymbol{\varepsilon}}_{t|t} = \mathrm{E}(\boldsymbol{\varepsilon}_t|\boldsymbol{\mathcal{Y}}_t, \boldsymbol{\mathcal{X}}_t, \boldsymbol{\mathcal{W}}_t)$, we have

$$\begin{aligned}\tilde{\boldsymbol{\varepsilon}}_{t|t} &= \mathbf{G}_t'\mathbf{F}_t^{*-1}(\boldsymbol{\nu}_t^* - \mathbf{V}_t\tilde{\boldsymbol{\beta}}_t) \\ &= \mathbf{G}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t.\end{aligned} \tag{11}$$

The real-time estimate of σ^2 is given by:

$$\hat{\sigma}_t^2 = \frac{1}{tN}\left[\sum_{i=1}^{t}\boldsymbol{\nu}_i\mathbf{F}_i^{-1}\mathbf{F}_i^*\mathbf{F}_i^{-1}\boldsymbol{\nu}_i\right].$$

Notice that the same real-time estimates as in (10) would be obtained from:

$$\begin{aligned}&\tilde{\boldsymbol{\alpha}}_{t|t}^* = \tilde{\boldsymbol{\alpha}}_{t|t-1}^* + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1}\boldsymbol{\nu}_t^*, \qquad \mathbf{A}_{t|t} = \mathbf{A}_{t|t-1} + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1}\mathbf{V}_t, \\ &\mathbf{P}_{t|t}^* = \mathbf{P}_{t|t-1}^* - \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\mathbf{F}_t^{*-1}\mathbf{Z}_t\mathbf{P}_{t|t-1}^*,\end{aligned} \tag{12}$$

setting

$$\tilde{\boldsymbol{\alpha}}_{t|t} = \tilde{\boldsymbol{\alpha}}_{t|t}^* - \mathbf{A}_{t|t}\tilde{\boldsymbol{\beta}}_t, \quad \mathbf{P}_{t|t} = \mathbf{P}_{t|t}^* + \mathbf{A}_{t|t}\mathbf{B}_t\mathbf{A}_{t|t}'. \tag{13}$$

In the particular case when $\mathbf{H}_t\mathbf{G}_t' = \mathbf{0}$, the prediction step for the state vector gives:

$$\begin{aligned}&\tilde{\boldsymbol{\alpha}}_{t+1|t}^* = \mathbf{T}_t\tilde{\boldsymbol{\alpha}}_{t|t}^*, \qquad \mathbf{A}_{t+1|t} = \mathbf{T}_t\mathbf{A}_{t|t} - \mathbf{W}_t, \\ &\mathbf{P}_{t+1|t}^* = \mathbf{T}_t\mathbf{P}_{t|t}^*\mathbf{T}_t' + \mathbf{H}_t\mathbf{H}_t'.\end{aligned} \tag{14}$$

3. The Data Cleaning Filter

3.1. M-estimator

To control for the effects of outliers, the AKF equations in Section 2 have to be modified. The starting point of our robustification procedure is the concept of an M-estimator. Let us refer, for simplicity, to the following univariate model: $y_t = m_t(\boldsymbol{\theta}) + \sigma_t u_t, t = 1, \ldots, n$, where the location $m_t(\boldsymbol{\theta})$ and the scale $\sigma_t(\boldsymbol{\theta})$ depend on a $(p \times 1)$ vector of unknown parameters. The vector $\boldsymbol{\theta}$ is estimated by minimizing the following criterion function, defined through a differentiable function ρ:

$$\min \sum_{t=1}^{n}\rho(u_t(\boldsymbol{\theta})), \tag{15}$$

where $u(\boldsymbol{\theta}) = (y_t - m_t(\boldsymbol{\theta}))/\sigma_t(\boldsymbol{\theta})$. The first order conditions derived from (15) are:

$$\sum_{t=1}^{n} \psi\left(u_t(\hat{\boldsymbol{\theta}})\right) \frac{\partial u_t}{\partial \hat{\boldsymbol{\theta}}_j} = 0, \qquad j = 1, ..., p. \tag{16}$$

Solution for $\hat{\boldsymbol{\theta}}$ obtained from (16) gives an M-estimator of $\boldsymbol{\theta}$. In (16), ψ is the first derivative of a ρ-function. Using a ψ-function is equivalent to scaling standardized residuals with a weight function w such that $w(u_t) = \psi(u_t)/u_t, t = 1, ..., n$. Any M-estimator defined by a differentiable M-function corresponds to an M-estimator defined by a ψ-function. The class of ψ-type M-estimators is actually broader, since the ψ-function is not required to be the partial derivative of any function of $\boldsymbol{\theta}$; see Hampel et al. (1986).

Note that if $\rho = -\log f$, with f being the standard normal density function ($u_t, t = 1, ..., n$, are assumed to have standard normal distribution), then the M-estimator $\hat{\boldsymbol{\theta}}$ is the ML estimator, which is equivalent to choosing $\rho(u_t) = u_t^2/2$, so that $\psi(u_t) = u_t$ and $w(u_t) = 1$ for $t = 1, ..., n$. Outliers in the data may result in large residuals that remain untreated in the case of ML (least squares) estimation, which may lead to biased estimates of $\boldsymbol{\theta}$. A remedy could be a more appropriate ψ-function that tries to diminish the effect of outliers by scaling large standardized residuals.[2] For redescending ψ-functions, $\psi(u) = 0$ for all $|u| \geq c$, with $0 < c < \infty$, so that the effect of high values of u is zero, whereas for monotone non-decreasing functions the effect is bounded by a constant d, i.e. $\psi(u) \leq d$, for $u > c > 0$. The corresponding M-estimators are called redescending and monotone M-estimators, respectively.

3.2. The Robust Augmented Kalman Filter

The approach based on scaling standardized residuals, that yields a robust M-estimator, can be transferred to our context as follows. Consider the factorization $\mathbf{F}_t = \mathbf{F}_t^{1/2}\mathbf{F}_t^{1/2'}$, which can be obtained, for instance from the Choleski decomposition of the matrix $\mathbf{F}_t$. Let $\mathbf{u}_t = \mathbf{F}_t^{-1/2}\boldsymbol{\nu}_t$ denote the vector of orthogonalized innovations, with $u_{it}, i = 1, \ldots, N$, being the individual elements of this vector. Further, let $\boldsymbol{\psi}(\mathbf{u}_t) = (\psi(u_{1i}), \ldots, \psi(u_{Nt}))'$ and $\boldsymbol{\Delta}_t = \text{diag}\,(w(u_{1i}), \ldots, w(u_{Nt}))$. Details on the specific form of the ψ-function considered in this paper are postponed to Section 3.5.

The robustification that we propose has the following logic. We define $\tilde{\mathbf{u}}_t = \boldsymbol{\Delta}_t\mathbf{u}_t$ and assume that $\tilde{\mathbf{u}}_t \sim \text{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$. This means that the orthogonalized innovations have

[2]The sensitivity of an M-estimator to outliers can be described by the so-called influence function. It measures the reaction of an estimator when the sample is contaminated by a small number of identical outliers. The influence function is proportional to a ψ-function evaluated at the outlier magnitude, which indicates that M-estimators derived from ψ-functions being less tolerant with respect to large residuals are less sensitive to outliers. For details, see Maronna et al. (2006).

the maintained Gaussian homoscedastic distribution only after rescaling them by $w(u_{it})$. In other words, the underlying innovation is distilled from the contaminated observable innovation u_{it}, by retaining only the fraction $w(u_{it})$ of it. Under this assumption, $\mathbf{u}_t \sim \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Delta}_t^{-2})$, which implies that the variance of the potentially outlying observations (those for which $w(u_{it}) < 1$) is greater than σ^2, and may even diverge to infinity, if $w(u_{it}) = 0$.

As a result, the underlying (clean) unobserved innovations, $\tilde{\boldsymbol{\nu}}_t = \mathbf{F}_t^{1/2}\tilde{\mathbf{u}}_t$, are IID $\mathrm{N}(\mathbf{0}, \mathbf{F}_t)$, whereas the observed contaminated innovations, after writing $\boldsymbol{\nu}_t = \mathbf{F}_t^{1/2}\boldsymbol{\Delta}_t^{-1}\mathbf{F}_t^{-1/2}\tilde{\boldsymbol{\nu}}_t$, have a normal distribution with zero mean and covariance matrix $\bar{\mathbf{F}}_t = \mathbf{F}_t^{1/2}\boldsymbol{\Delta}_t^{-2}\mathbf{F}_t^{1/2\prime}$. In other words, the innovations $\boldsymbol{\nu}_t$ computed by the AKF are contaminated, while $\mathbf{F}_t$ is the covariance matrix of clean innovations, rather than that of $\boldsymbol{\nu}_t$.

Our proposed robust AKF replaces $\mathbf{F}_t^{-1}$ by $\bar{\mathbf{F}}_t^{-1} = \mathbf{F}_t^{-1/2\prime}\boldsymbol{\Delta}_t^2\mathbf{F}_t^{-1/2}$ in the updating equations for the state and the regression effects. In particular, after running (3) and (4), $t = k+1, \ldots, n$, we compute

$$\begin{array}{rcl} \tilde{\boldsymbol{\beta}}_t &=& \tilde{\boldsymbol{\beta}}_{t-1} + \mathbf{B}_{t-1}\mathbf{V}_t'\bar{\mathbf{F}}_t^{-1}\boldsymbol{\nu}_t, \\ &=& \tilde{\boldsymbol{\beta}}_{t-1} + \mathbf{B}_{t-1}\mathbf{V}_t'\mathbf{F}_t^{-1/2\prime}\boldsymbol{\Delta}_t^2\mathbf{F}_t^{-1/2}\boldsymbol{\nu}_t, \\ \mathbf{B}_t &=& \mathbf{B}_{t-1} - \mathbf{B}_{t-1}\mathbf{V}_t'\bar{\mathbf{F}}_t^{-1}\mathbf{V}_t\mathbf{B}_{t-1}, \\ &=& \mathbf{B}_{t-1} - \mathbf{B}_{t-1}\mathbf{V}_t'\mathbf{F}_t^{-1/2\prime}\boldsymbol{\Delta}_t^2\mathbf{F}_t^{-1/2}\mathbf{V}_t\mathbf{B}_{t-1}, \end{array} \tag{17}$$

Note that $\boldsymbol{\Delta}_t = \mathbf{I}_N$ yields the usual updated estimate, whereas if $\boldsymbol{\Delta}_t = \mathbf{0}$, $\tilde{\boldsymbol{\beta}}_t = \tilde{\boldsymbol{\beta}}_{t-1}$ and $\mathbf{B}_t = \mathbf{B}_{t-1}$, so that the updating of the inferences does not take place.

Outlier contamination also affects the observed conditional innovations $\boldsymbol{\nu}_t^*$. As a matter of fact, writing:

$$\begin{array}{rcl} \boldsymbol{\nu}_t^* &=& \boldsymbol{\nu}_t + \mathbf{V}_t\tilde{\boldsymbol{\beta}}_{t-1} \\ &=& \mathbf{F}_t^{1/2}\boldsymbol{\Delta}_t^{-1}\mathbf{F}_t^{-1/2}\tilde{\boldsymbol{\nu}}_t + \mathbf{V}_t\tilde{\boldsymbol{\beta}}_{t-1}, \end{array}$$

we have that the covariance matrix of the innovations is inflated by the presence of outliers. Hence, we can define the underlying conditional innovations $\tilde{\boldsymbol{\nu}}_t^* = \tilde{\boldsymbol{\nu}}_t + \mathbf{V}_t\tilde{\boldsymbol{\beta}}_{t-1}$, $\tilde{\boldsymbol{\nu}}_t^* \sim$ iid $\mathrm{N}(\mathbf{0}, \mathbf{F}_t^*)$, $\mathbf{F}_t^* = \mathbf{F}_t - \mathbf{V}_t\mathbf{B}_{t-1}\mathbf{V}_t'$, so that the observed contaminated innovations $\boldsymbol{\nu}_t^*$ have covariance matrix $\bar{\mathbf{F}}_t^* = \bar{\mathbf{F}}_t - \mathbf{V}_t\mathbf{B}_{t-1}\mathbf{V}_t'$. The robustified estimates are thus obtained by replacing $\mathbf{F}_t^*$ by $\bar{\mathbf{F}}_t^*$, whose inverse is:

$$\bar{\mathbf{F}}_t^{*-1} = \bar{\mathbf{F}}_t^{-1} + \bar{\mathbf{F}}_t^{-1}\mathbf{V}_t\left(\mathbf{B}_{t-1} - \mathbf{V}_t'\bar{\mathbf{F}}_t^{-1}\mathbf{V}_t\right)^{-1}\mathbf{V}_t'\bar{\mathbf{F}}_t^{-1}.$$

The robustified real-time estimates of the state vector are:

$$\begin{array}{l} \tilde{\boldsymbol{\alpha}}_{t|t} = \tilde{\boldsymbol{\alpha}}_{t|t-1}^* - \mathbf{A}_{t|t-1}\tilde{\boldsymbol{\beta}}_t + \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\bar{\mathbf{F}}_t^{-1}\boldsymbol{\nu}_t, \\ \mathbf{P}_{t|t} = \mathbf{P}_{t|t-1}^* - \mathbf{P}_{t|t-1}^*\mathbf{Z}_t'\bar{\mathbf{F}}_t^{*-1}\mathbf{Z}_t\mathbf{P}_{t|t-1}^* + \mathbf{A}_{t|t}\mathbf{B}_{t|t}\mathbf{A}_{t|t}', \end{array} \tag{18}$$

Finally, the robustified real-time estimate of the disturbance vector is:

$$\tilde{\boldsymbol{\varepsilon}}_{t|t} = \mathbf{G}_t' \bar{\mathbf{F}}_t^{-1} \boldsymbol{\nu}_t. \tag{19}$$

If $\boldsymbol{\Delta}_t = \mathbf{0}$, it can be easily seen from (18) that $\tilde{\boldsymbol{\alpha}}_{t|t} = \tilde{\boldsymbol{\alpha}}_{t|t-1}$. Moreover, $\tilde{\boldsymbol{\alpha}}_{t+1|t}$ becomes a two-step-prediction in this case, which becomes evident from writing $\tilde{\boldsymbol{\alpha}}_{t+1|t} = \mathbf{T}_t \tilde{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{W}_t \tilde{\boldsymbol{\beta}}_t + \mathbf{H}_t \tilde{\boldsymbol{\varepsilon}}_{t|t}$, and replacing $\tilde{\boldsymbol{\beta}}_t$ by (17), and $\tilde{\boldsymbol{\varepsilon}}_{t|t}$ by (19).

Also, denoting $\bar{\mathbf{K}}_t^* = (\mathbf{T}_t \mathbf{P}_{t|t-1} \mathbf{Z}_t' + \mathbf{H}_t \mathbf{G}_t') \bar{\mathbf{F}}_t^{*-1}$, the prediction equations become

$$\begin{aligned}
\tilde{\boldsymbol{\alpha}}_{t+1|t}^* &= \mathbf{T}_t \tilde{\boldsymbol{\alpha}}_{t|t-1}^* + \bar{\mathbf{K}}_t^* \boldsymbol{\nu}_t^*, \\
\mathbf{A}_{t+1|t} &= \mathbf{T}_t \mathbf{A}_{t|t-1} - \mathbf{W}_t + \bar{\mathbf{K}}_t^* \mathbf{V}_t, \\
\mathbf{P}_{t+1|t}^* &= \mathbf{T}_t \mathbf{P}_{t|t-1}^* \mathbf{T}_t' + \mathbf{H}_t \mathbf{H}_t' - \bar{\mathbf{K}}_t^* \bar{\mathbf{F}}_t^* \bar{\mathbf{K}}_t^{*\prime}.
\end{aligned} \tag{20}$$

The robust AKF replaces $\mathbf{y}_t$ with

$$\mathbf{y}_t^\dagger = \mathbf{Z}_t \tilde{\boldsymbol{\alpha}}_{t|t} + \mathbf{X}_t \tilde{\boldsymbol{\beta}}_t + \mathbf{G}_t \tilde{\boldsymbol{\varepsilon}}_{t|t},$$

where $\tilde{\boldsymbol{\alpha}}_{t|t}$, $\tilde{\boldsymbol{\beta}}_t$, and $\tilde{\boldsymbol{\varepsilon}}_{t|t}$ are as in (18), (17), and (19), respectively. The quantity $\mathbf{y}_t^\dagger$ is equal to $\mathbf{y}_t$ if $\boldsymbol{\Delta}_t = \mathbf{I}_N$; if, however, $\boldsymbol{\Delta}_t$ tends towards zero, the robust AKF shrinks an outlying observation towards the one-step-ahead prediction $\tilde{\mathbf{y}}_{t|t-1} = \mathbf{Z}_t \tilde{\boldsymbol{\alpha}}_{t|t-1} + \mathbf{X}_t \tilde{\boldsymbol{\beta}}_{t-1}$. The sequence $\{\mathbf{y}_t^\dagger\}$, $t = 1, \ldots, n$, represents a cleaned data set which can be used for robust parameter estimation.

3.3. Discussion

It is possible to think of two alternative strategies for adjusting the AKF. The first one aims at correcting $\boldsymbol{\nu}_t$, so that the corrected innovations, $\tilde{\boldsymbol{\nu}}_t$ have covariance matrix $\mathbf{F}_t$, up to a scale factor. The second, which is the one we propose, relies on adjusting the covariance matrix $\mathbf{F}_t$ to $\bar{\mathbf{F}}_t$, which is the covariance matrix of the observed innovations. We decide to follow the latter strategy, since an outlier should not affect the inferences on the predictive distribution of the states and the observables. In fact, our chosen strategy implies that in the presence of a potential outlier, in the extreme case when $\boldsymbol{\Delta}_t = \mathbf{0}$, no updating takes place for $\tilde{\boldsymbol{\beta}}_t$, $\boldsymbol{\alpha}_{t|t}$, as well as their respective covariance matrices $\mathbf{B}_t$ and $\mathbf{P}_{t|t}$. An outlying observation is uninformative about the current state and its one-step ahead prediction. In contrast, the first strategy would involve updating covariance matrices even in an extreme situation of such a big outlier so that the innovations are reduced to zero. This can be directly seen for $\mathbf{B}_t$ in eq. (8); as $\mathbf{F}_t$ is never adjusted towards zero, $\mathbf{B}_{t-1}$ is always being updated to $\mathbf{B}_t$.

For an alternative strategy, which does not adjust the recursions for the covariance

matrices, see Ruckdeschel et al. (2014). To be precise, in their setup without regression effects and with an ordinary Kalman filter (i.e., when $\tilde{\boldsymbol{\alpha}}_{t|t} = \tilde{\boldsymbol{\alpha}}^*_{t|t}$), as an adjustment of $\tilde{\boldsymbol{\alpha}}^*_{t|t}$ the authors propose clipping all coordinates rather than adjusting $\boldsymbol{\nu}^*_t$ only, i.e.:

$$\tilde{\boldsymbol{\alpha}}^*_{t|t} = \tilde{\boldsymbol{\alpha}}^*_{t|t-1} + \Omega(\mathbf{P}^*_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{*-1}\boldsymbol{\nu}^*_t),$$

where $\Omega(\mathbf{x})$ represents a scaling function whose values depend on $|\mathbf{x}|$, length of $\mathbf{x}$ given by, e.g., the Euclidian norm. For a Huber function (which will be introduced in the next subsection) chosen by the authors, $\Omega(\mathbf{x}) = \mathbf{x}\min(1, c/|\mathbf{x}|)$, with c being a tuning parameter. In our setup, this approach would be based on the following modification of the real-time estimates $\tilde{\boldsymbol{\alpha}}_{t|t}$ and $\tilde{\boldsymbol{\beta}}_t$ (apart from an appropriate modification of the prediction step for $\tilde{\boldsymbol{\alpha}}^*_{t|t}$):

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_t &= \tilde{\boldsymbol{\beta}}_{t-1} + \Omega(\mathbf{B}_{t-1}\mathbf{V}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t),\\ \tilde{\boldsymbol{\alpha}}_{t|t} &= \tilde{\boldsymbol{\alpha}}^*_{t|t-1} - \mathbf{A}_{t|t-1}\tilde{\boldsymbol{\beta}}_t + \Omega(\mathbf{P}^*_{t|t-1}\mathbf{Z}_t'\mathbf{F}_t^{-1}\boldsymbol{\nu}_t).\end{aligned}$$

As a final point, it is to be noted that, irrespective of the outlier type, the goal of our robustification procedure is to recover the nominal behavior of the series, i.e. the behavior in the absence of any outliers. Alternative goals depending on the outlier type can lead to alternative robustification approaches. For example, for outliers which change the series persistently, like the innovation outliers, a legitimate goal would be to follow the new behavior of the series after it has been affected by an outlier. Such an aim is reflected in the approach proposed by Ruckdeschel et al. (2014).

3.4. Estimation of the model hyperparameters

The ML estimator of the parameters is not robust to outliers. A robust M–type estimates can be obtained by the following procedure:

1. Compute the ML estimates of $\boldsymbol{\theta}$ and obtain a robust scale estimate by replacing (6) by the median absolute deviation of the standardized innovations:

$$\left[\text{med}\left(|u_{it} - \text{med}(u_{it})|\right)/0.6745\right]^2,$$

   where $\text{med}(\cdot)$ is the median of the distribution.
2. Run the robust AKF of Section 3.2 to obtain a clean series.
3. Estimate the parameters on the clean series by ML.

Steps 2–3 may be iterated until the robust AKF coincides with the AKF and no further corrections to the series are made.

3.5. The ψ-function

The ψ-function is an essential building block of the robust AKF. In our applications, we shall use the Huber function which, for a variable x, is given by

$$\psi(x) = \begin{cases} x, & \text{if} \quad |x| \leq c \\ d = c \ \text{sign}(x), & \text{if} \quad |x| > c \end{cases}$$

The Huber function belongs to the class of monotone non-decreasing ψ-functions and is the most-widely used ψ-function. The tuning constant c regulates the trade-off between the so-called breakdown point and the efficiency of the estimator. The breakdown point is a measure of robustness of an estimator as it gives the fraction of bad data the estimator can tolerate before giving results towards the boundary of the parameter space. Lower values of c increase the breakdown point but reduce efficiency. We set $c = 1.345$, which guarantees 95% efficiency when sampling from the normal distribution. The values of the Huber function, $\psi(x)$, and the corresponding values of the weight function, $w(x) = \psi(x)/x$, using $c = 1.345$, are depicted in Figure 1.

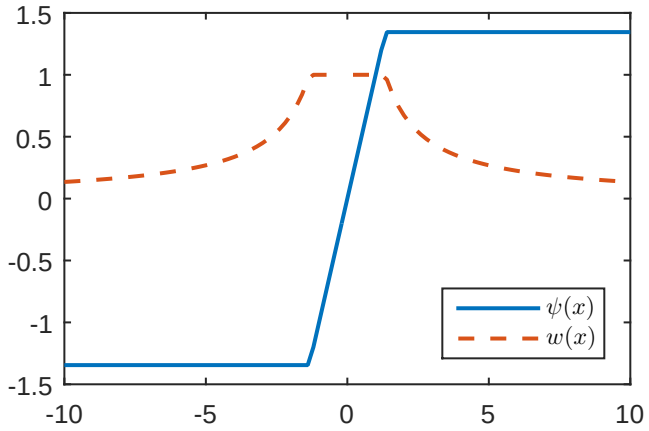


Figure 1: Values of the Huber function ψ and the corresponding weight function w for $c = 1.345$.

4. Modeling Framework: the Basic Structural Model

After the robust AKF has been introduced for a general class of linear Markovian models, in the following we review the BSM, which will serve as the common modeling framework in the application part of the article.

The BSM postulates an additive and orthogonal decomposition of a time series into unobserved components representing the trend, seasonality and the irregular component.

If y_t denotes a time series observed at $t = 1, 2, \ldots, n$, the decomposition can be written as follows:

$$y_t = \mu_t + \gamma_t + \epsilon_t, \quad t = 1, \ldots, n, \tag{21}$$

where μ_t is the trend component, γ_t is the seasonal component, and $\epsilon_t \sim$ IID $\mathrm{N}(0, \sigma_\epsilon^2)$ is the irregular component.[3]

The trend component has a local linear representation:

$$\begin{array}{rcll} \mu_{t+1} & = & \mu_t + & \rho_t + \eta_t \\ \rho_{t+1} & = & & \rho_t + \zeta_t, \end{array} \tag{22}$$

where η_t and ζ_t are mutually and serially uncorrelated normally distributed random shocks with zero mean and variances σ_η^2 and σ_ζ^2, respectively.

The seasonal component can be modeled as a combination of six stochastic cycles whose common variance is σ_ω^2. The single stochastic cycles have a trigonometric representation and are defined at the seasonal frequencies $\lambda_j = 2\pi j/12$, $j = 1, \ldots, 6$. The parameter λ_1 denotes the fundamental frequency (corresponding to a period of 12 monthly observations) and the remaining ones represent the five harmonics (corresponding to periods of 6 months, i.e. two cycles in a year, 4 months, i.e. three cycles in a year, 3 months, i.e. four cycles in a year, 2.4, i.e. five cycles in a year, and 2 months):

$$\gamma_t = \sum_{j=1}^{6} \gamma_{jt}, \qquad \begin{bmatrix} \gamma_{j,t+1} \\ \gamma_{j,t+1}^* \end{bmatrix} = \begin{bmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{bmatrix} \begin{bmatrix} \gamma_{j,t} \\ \gamma_{j,t}^* \end{bmatrix} + \begin{bmatrix} \omega_{j,t} \\ \omega_{j,t}^* \end{bmatrix}, j = 1, \ldots, 5, \tag{23}$$

and $\gamma_{6,t+1} = -\gamma_{6t} + \omega_{6t}$. The disturbances ω_{jt} and ω_{jt}^* are normally and independently distributed with common variance σ_ω^2 for $j = 1, \ldots, 5$, whereas $\mathrm{Var}(\omega_{6t}) = 0.5\sigma_\omega^2$.

The state space representation of the BSM has $m = 13$ state components, $\boldsymbol{\alpha}_t = [\mu_t, \rho_t, \gamma_{1t}, \gamma_{1t}^*, \ldots, \gamma_{6t}]'$, and disturbances

$$\boldsymbol{\varepsilon}_t = \sigma \left[\frac{\epsilon_t}{\sigma_\epsilon}, \frac{\eta_t}{\sigma_\eta}, \frac{\zeta}{\sigma_\zeta}, \frac{\omega_{1t}}{\sigma_\omega}, \ldots, \frac{\omega_{6t}}{\sigma_\omega} \right]'.$$

The system matrices are time invariant, $\mathbf{Z}_t = \mathbf{Z}, \mathbf{G}_t = \mathbf{G}, \mathbf{T}_t = \mathbf{T}, \mathbf{H}_t = \mathbf{H}$, and

$$\mathbf{Z} = [1, 0, 1, 0, \ldots, 1], G = \left[\frac{\sigma_\epsilon}{\sigma}, 0, \ldots, 0 \right],$$

[3]Eq. (21) can additionally include regressors that account for any known interventions as well as calendar effects which are, apart from outlier effects, typically also removed during seasonal adjustment.

$$\mathbf{T} = \begin{bmatrix} 1 & 1 & 0 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & \cos\lambda 1 & \sin\lambda_1 & \dots & \dots & 0 \\ 0 & 0 & -\sin\lambda 1 & \cos\lambda_1 & \dots & \dots & 0 \\ \vdots & \vdots & \dots & \dots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \dots & -1 \end{bmatrix}, \mathbf{H} = \frac{1}{\sigma} \begin{bmatrix} 0 & \sigma_\eta & 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & \sigma_\omega & 0 & \dots & \dots & 0 \\ 0 & 0 & 0 & \sigma_\omega & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & \ddots & \dots & 0 \\ \vdots & \vdots & \dots & \dots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \dots & \sqrt{0.5}\sigma_\omega \end{bmatrix}.$$

Moreover, it holds that $\mathbf{W}_t = \mathbf{0}$, $t = 1, \dots, n$. The scale parameter σ^2 is set equal to one of the variance parameters, and we adopt $\sigma^2 = \sigma_\epsilon^2$ as a default. The initial conditions are set as follows: $\tilde{\boldsymbol{\alpha}}^*_{1|0} = \mathbf{0}$, $\mathbf{W}_0 = \mathbf{T}$ ($\boldsymbol{\beta} = \tilde{\boldsymbol{\alpha}}_{1|0}$), and $\mathbf{H}_0 = \mathbf{H}$.

In the next two sections, we illustrate an application of the robust AKF to structural time series models using the BSM. First, in Section 5, we conduct a Monte Carlo experiment to assess the gains in the precision of the parameter estimates when additive or innovation outliers are present in the simulated data. As detection of outliers of this kind may also have consequences for forecasting, in the second step, in Section 6, we illustrate an application of the robust AKF in the context of forecasting. For this purpose, we use a large set of real data contaminated by outliers.

5. A Monte Carlo Experiment

5.1. Design of the experiment

We generate time series of length $n = 144$ observations (12 years of monthly data) using the data generating process (DGP) given by the BSM in eq. (23).[4] We distinguish between 5 scenarios regarding variance parameters regulating the DGP:

- the benchmark scenario with $\sigma_\epsilon^2 = 1, \sigma_\eta^2 = 0.08, \sigma_\zeta^2 = 0.0001, \sigma_\omega^2 = 0.05$. As the irregular variance is set equal to 1, the remaining parameters are interpreted as signal to noise ratios. The benchmark DGP is chosen on the basis of our experience in fitting the BSM to industrial production and turnover time series. Note that the values of σ_ϵ^2 and σ_ζ^2 remain unchanged across different scenarios.

- a stable trend–stable seasonal scenario (labeled sT–sS) with $\sigma_\eta^2 = 0.00008$ and $\sigma_\omega^2 = 0.00005$

- a stable trend–unstable seasonal scenario (sT–uS) with $\sigma_\eta^2 = 0.00008$ and $\sigma_\omega^2 = 0.5$

[4]The initial values for the components are: $\mu_0 = 91.06$, $\rho_0 = 0.00015$, $[\gamma_{1,0}, \gamma^*_{1,0}] = [-0.381; 4.1483]$, $[\gamma_{2,0}; \gamma^*_{2,0}] = [-6.863; -4.00136]$, $[\gamma_{3,0}; \gamma^*_{3,0}] = [-3.41264; 9.99139]$, $[\gamma_{4,0}; \gamma^*_{4,0}] = [2.032516; -5.47096]$, $[\gamma_{5,0}; \gamma^*_{5,0}] = [-6.65170; 2.93962]$, $\gamma_{6,0} = 5.88545$. These values correspond to the values of the respective smoothed components of the Italian industrial production series in 1995.

- an unstable trend–stable seasonal scenario (uT–sS) with $\sigma_\eta^2 = 0.8$ and $\sigma_\omega^2 = 0.00005$
- an unstable trend–unstable seasonal scenario (uT–uS) with $\sigma_\eta^2 = 0.8$ and $\sigma_\omega^2 = 0.5$

The simulated series are contaminated with randomly located and sized outliers of a particular type. The considered outlier types are: additive outliers (AOs), occurring either individually or in a patch, and innovation outliers (IOs). As a result of outlier contamination, we observe $y_t = y_t^\ddagger + \xi_t$, where $y_t^\ddagger$ is generated according to the BSM and ξ_t represents the outlier effect depending on the outlier type.

The AO generating model is $z_t\,\delta\,I_t$, where δ denotes the reference size, z_t is IID N(0,1), and I_t is an IID Bernoulli random variable with success probability $p = 0.02$. The value of δ and the consequences of having a random z_t will be discussed later. As for the patch of AOs, we allow only for a single patch of k consecutive outliers located at a random time τ. More specifically, we draw k from a discrete uniform distribution with support $\{3,\ 4,\ \ldots,\ 12\}$; the location τ is drawn at random from a uniform distribution with support $\{1, 2, \ldots, n-k+1\}$; setting $I_t(\tau, k) = 1$ for $t = \tau, \ldots, \tau + k - 1$, the AO patch is generated by $z_t\,\delta\,I_t(\tau, k)$. In the case of individual AOs, it holds that $\xi_t = \delta z_t I_t$ and for a patch of AOs ξ_t is given by $\xi_t = \delta z_t I_t(\tau, k)$. This means that the outlier signature, i.e. the influence of an outlier occurring at a particular time point on the current and future observations, coincides in both AO cases with the outlier magnitude $z_t\delta$. Examples of simulated series featuring single random AOs and a random AO patch as well as the corresponding outlier effects are presented in Figure 2, panel a)–b) (single random AOs), and panel c)–d) (random AO patch).

In the IO case, for the location τ at which the outlier occurs, we define the dummy variable taking values

$$D_t(\tau) = \begin{cases} 0, & t = 1, \ldots, \tau - 1 \\ 1, & t = \tau \\ \mathbf{Z}\mathbf{T}^{t-\tau-1}\mathbf{K}^*, & t = \tau + 1, \ldots, n, \end{cases}$$

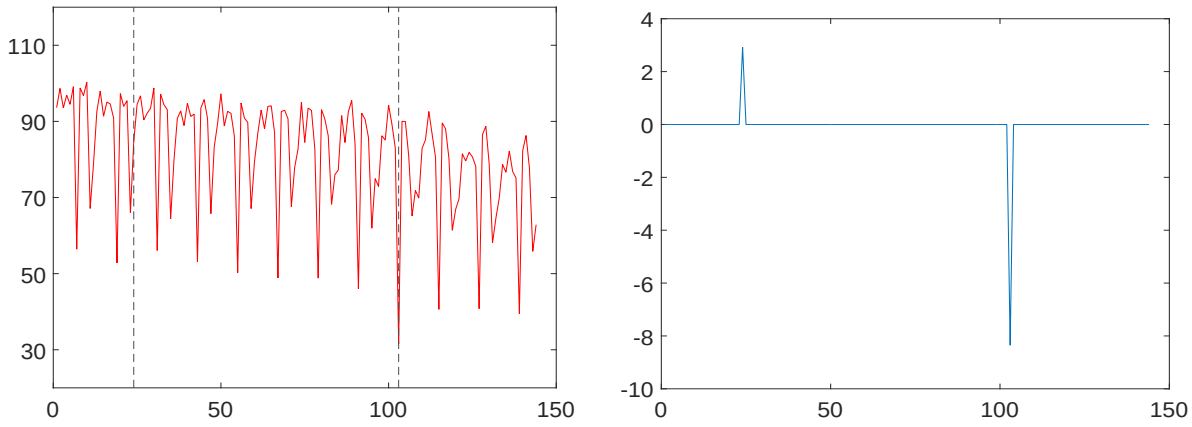
where $\mathbf{K}^*$ denotes the Kalman gain in the steady state. The outlier signature is thus given by the impulse–response function derived from the innovation form of the state space model (1). The occurrence of outliers is, similarly as in the AO case, governed by an IID Bernoulli random variable I_t with success probability $p = 0.02$. If $I_t = 1$ at locations $t = \tau_j, j = 1, \ldots, J$, then the outlier effect at each time point t is given by $\xi_t = \sum_{j=1}^{J} z_j\,\delta\,D_t(\tau_j)$, where z_j are IID standard normal draws. Panels e)–f) of Figure 2 illustrate an example of a simulated series that is affected by single random IOs, and the effect of such random IOs.

The reference size δ is expressed by $\delta = 7{\cdot}\text{PESD}$ with $\text{PESD} = \sigma F^{1/2}$ denoting the prediction error standard deviation, which is obtained from the innovations form of the model in the steady state ($F = \lim_{t\to\infty} F_t$). The PESD increases with σ_η^2 and σ_ω^2 and attains the highest value in the uT–uS scenario. Hence, tying δ to the PESD accounts for the difficulty of detecting outliers in the case of a high overall variation. A detailed discussion of the choice of the outlier magnitude for structural time series models is provided by Marczak and Proietti (2016), who consider the same settings regarding the model for simulations (BSM) and the values for the variance parameters. In addition to the reference size $\delta = 7{\cdot}\text{PESD}$, also chosen as a reference size in Marczak and Proietti (2016), we consider $\delta = 14{\cdot}\text{PESD}$. Increasing the magnitude of δ is motivated by the fact that, in contrast to Marczak and Proietti (2016), the final outlier size is not given by δ but is obtained by scaling δ with a standard normal variable z_t. Since the probability that z_t takes on values between -1 and 1 is 68.27%, the final outlier size is in most of the cases smaller than δ. A higher δ is thus supposed to countervail low values of z_t. It is to be noted that instead of scaling δ with a random number, we could have examined different values of δ obtained by scaling PESD with a range of factors. However, our setting is a more realistic one as it allows for different sizes of outliers affecting a particular series whereas the alternative setting would imply the same deterministic magnitude.

Taking into account the settings described above, for each combination of the variance parameters and outlier types a simulation experiment is conducted to evaluate the performance of the robust AKF. Every single experiment consists of the following steps:[5]

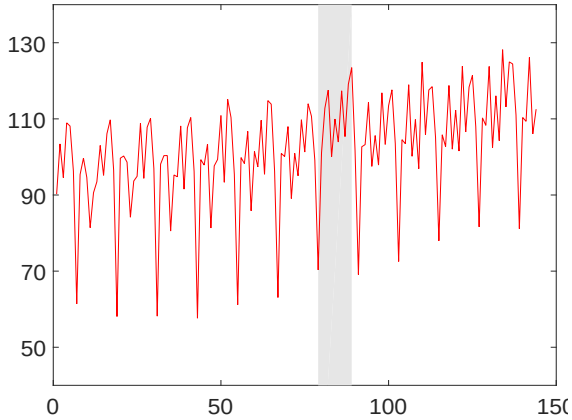
1. Obtain series contaminated with outliers using the BSM and an outlier generating process.
2. Fit the BSM to the simulated series and put the BSM into the state space form (1).
3. Run the ordinary AKF to the simulated series and the state space model and, by maximizing the likelihood function in eq. (5), obtain the ML estimates of the variance parameters: $\hat{\sigma}_\epsilon^2, \hat{\sigma}_\eta^2, \hat{\sigma}_\zeta^2, \hat{\sigma}_\omega^2$.
4. Apply the procedure described in Section 3.4 to obtain the robust estimates of the parameters: $\tilde{\sigma}_\epsilon^2, \tilde{\sigma}_\eta^2, \tilde{\sigma}_\zeta^2, \tilde{\sigma}_\omega^2$.
5. After 1000 replications of steps 1–4, compute the relative efficiency corresponding to each of the variance parameters, given by the ratio of the mean square error (MSE)

[5] All computations are performed with Matlab R2015a. We also experimented with different values of δ and, as an alternative to the Huber function, we also investigated redescending functions, which exhibit higher resistance to large outliers than monotone functions. More specifically, we considered the Cauchy and the Welsch functions. Both specifications show, in general, satisfactory results, qualitatively comparable to those obtained with the Huber function. These further outcomes of the simulation experiment corresponding to the alternative ψ-functions are available in the supplementary material to this article.

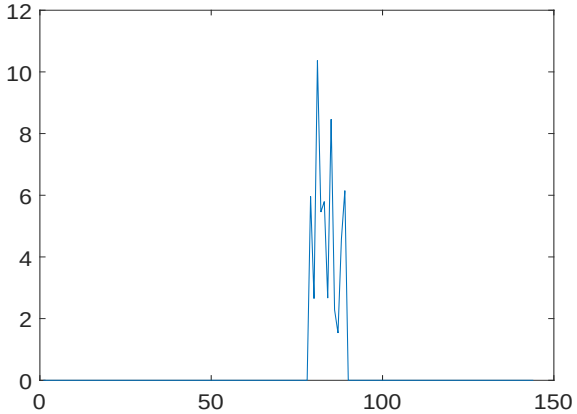


(a) Series with random AOs

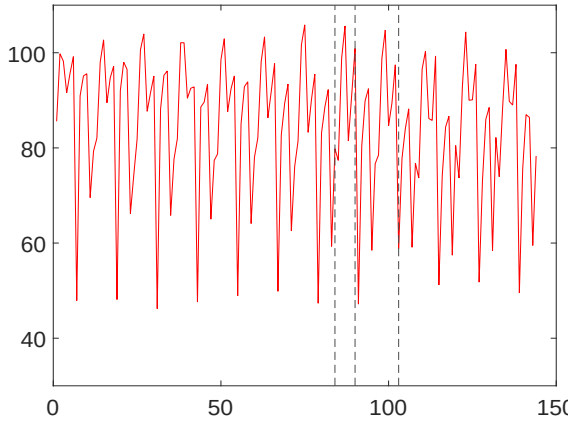
(b) Effects of random AOs



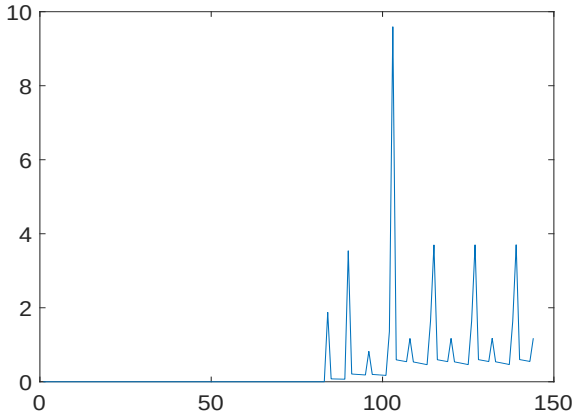
(c) Series with a random AO patch



(d) Effects of a random AO patch



(e) Series with random IOs



(f) Effects of random IOs

Figure 2: Examples of simulated series affected by random AOs, a random AO patch, and random IOs, respectively (left panel). Location of random outliers of a particular type is indicated by: vertical dashed lines in a) (AOs) and e) (IOs), and a shaded vertical band in c) (AO patch). The corresponding outlier effects are depicted in the right panel. The contaminated series are generated using the benchmark setting; the outlier size is set relative to $\delta = 7 \cdot$ PESD.

of the ML estimates to the MSE of the robust estimates.

5.2. Results

Table 1 summarizes the simulation results for series contaminated by AOs for the five different parameter scenarios outlined above, and two different values of δ regulating the outlier size. In particular, the table displays the relative efficiency of the ML estimator compared to that of the robust estimator of the variance parameters, as measured by the ratio of the respective MSE values. It is evident that for all parameters as well as both values of δ and all variance combinations, the robust estimator is more efficient than the ML estimator. Only for the seasonal variance, σ_ω^2, the efficiency ratio is in two of the ten considered cases below one. Increasing the reference size from the benchmark value to 14·PESD leads, however, to a considerable improvement of the results for the most variance parameters. Independently of δ, for the irregular variance, σ_ϵ^2, a substantial

Table 1: Series contaminated by random AOs: MSE ratio of the ML estimator to that of the robust estimator.

	Benchmark: $\delta = 7$·PESD					$\delta = 14$·PESD				
	Benchmark	sT–sS	uT–sS	sT–uS	uT–uS	Benchmark	sT–sS	uT–sS	sT–uS	uT–uS
σ_ϵ^2	12.17	6.89	13.57	8.87	12.88	11.00	44.24	19.72	10.73	11.68
σ_η^2	1.50	6.91	12.16	18.12	1.03	2.15	29.63	58.87	33.89	6.65
σ_ζ^2	3.83	1.19	10.01	1.85	3.99	4.65	1.19	20.07	3.70	3.56
σ_ω^2	0.63	7.85	27.09	1.76	0.40	2.03	17.40	39.18	4.80	1.73

efficiency gain is visible in all scenarios. In the benchmark case as well as in the uT–uS case, the gain for σ_ϵ^2 is the highest compared to other variance parameters. As regards the variance of the seasonal component, σ_ω^2, the biggest gain is attained when the seasonal component is stable. As for the variance of the level component, σ_η^2, the robust estimator performs better in terms of efficiency if either the level or the seasonal component is stable. In general, it can be concluded that the highest accuracy of the robust estimator is achieved if the seasonal component is stable, which means that in such a case it is easier to decouple the effect of outliers from the variation of a particular component. The results for the benchmark scenario confirm the superiority of the robust estimator.

The evaluation results for a random patch of AOs are reported in Table 2. As in the previous case, the robust estimator is in general more efficient than the ML estimator, especially if the reference size δ is doubled. Likewise, it holds that the efficiency gain for σ_ϵ^2 is the highest relative to other components' variances in the benchmark and in the uT–uS scenarios. For σ_ω^2, the robust estimator turns out to be less efficient in three out of five scenarios for the benchmark size of δ. However, it improves in efficiency relative

to the ML estimator if the reference size is larger. Moreover, it is clearly more efficient if the seasonal component is stable.

Table 2: Series contaminated by a random patch of AOs: MSE ratio of the ML estimator to that of the robust estimator.

	Benchmark: $\delta = 7$·PESD					$\delta = 14$·PESD				
	Benchmark	sT–sS	uT–sS	sT–uS	uT–uS	Benchmark	sT–sS	uT–sS	sT–uS	uT–uS
σ_ϵ^2	23.95	4.13	11.61	20.76	24.40	41.73	37.72	60.53	41.10	35.50
σ_η^2	14.05	9.23	18.41	6.21	9.65	11.23	26.25	60.41	15.29	14.84
σ_ζ^2	7.29	12.63	8.79	1.82	7.54	5.27	21.04	30.04	34.96	6.09
σ_ω^2	0.29	19.10	20.69	0.22	0.23	0.77	46.35	60.38	0.85	1.12

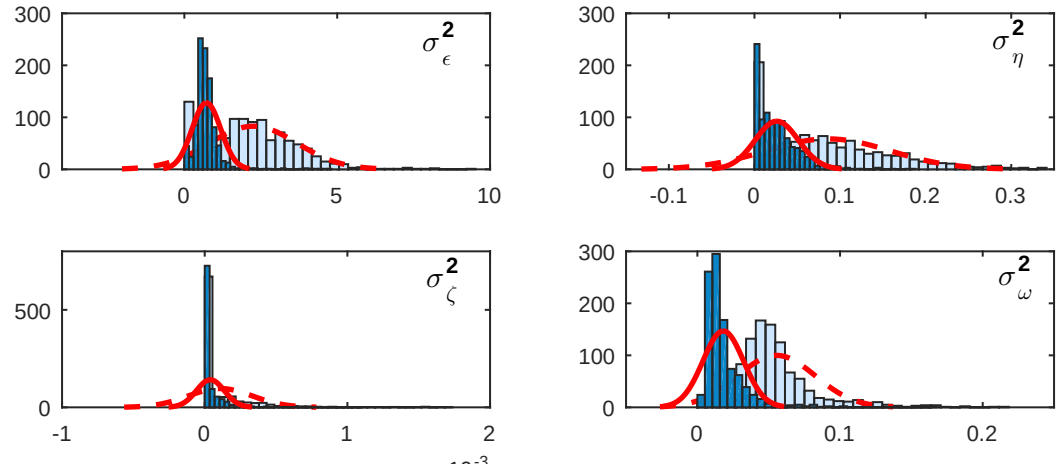
If the series are contaminated with random IOs (see Table 3), the robust estimator still remains more efficient than the ML estimator. For the variances of the level and the seasonal component, σ_η^2 and σ_ω^2, respectively, the efficiency gain is the highest if the seasonal component is stable, which resembles the outcomes in the case of individual AOs and AO patch.

Table 3: Series contaminated by random IOs: MSE ratio of the ML estimator to that of the robust estimator.

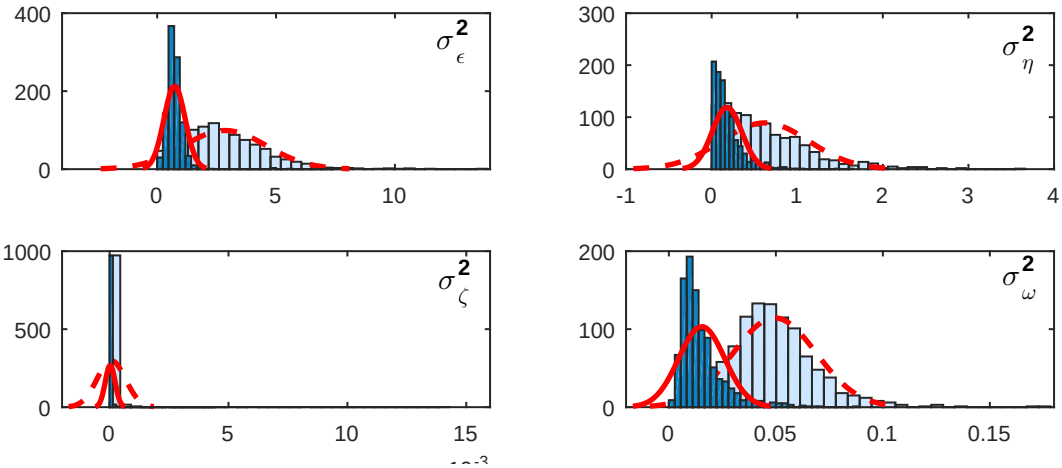
	Benchmark: $\delta = 7$·PESD					$\delta = 14$·PESD				
	Benchmark	sT–sS	uT–sS	sT–uS	uT–uS	Benchmark	sT–sS	uT–sS	sT–uS	uT–uS
σ_ϵ^2	8.10	8.41	3.96	2.32	10.32	12.67	64.42	29.54	2.44	18.34
σ_η^2	1.50	9.13	12.87	4.54	4.19	3.09	54.20	49.23	6.81	16.78
σ_ζ^2	4.00	1.17	8.43	1.58	3.69	4.81	1.48	14.34	3.52	9.30
σ_ω^2	0.59	6.59	19.49	2.94	3.50	1.72	21.76	43.78	2.28	37.22

The comparison of the distribution of the ML and robust estimates of the BSM parameters is presented in Figure 3 for the benchmark scenario. The plot is divided into three panels, according to the outlier type (AOs, AO patch and IOs). The light blue bars (histogram) and the dashed red line (density) correspond to the ML estimates, whereas the dark blue bars (histogram) and the solid red line (density) refer to the robust estimates. It can be observed that for all three types of outlier contamination the distribution of the ML estimates of all variance parameters is more dispersed than the distribution of the robust estimates.

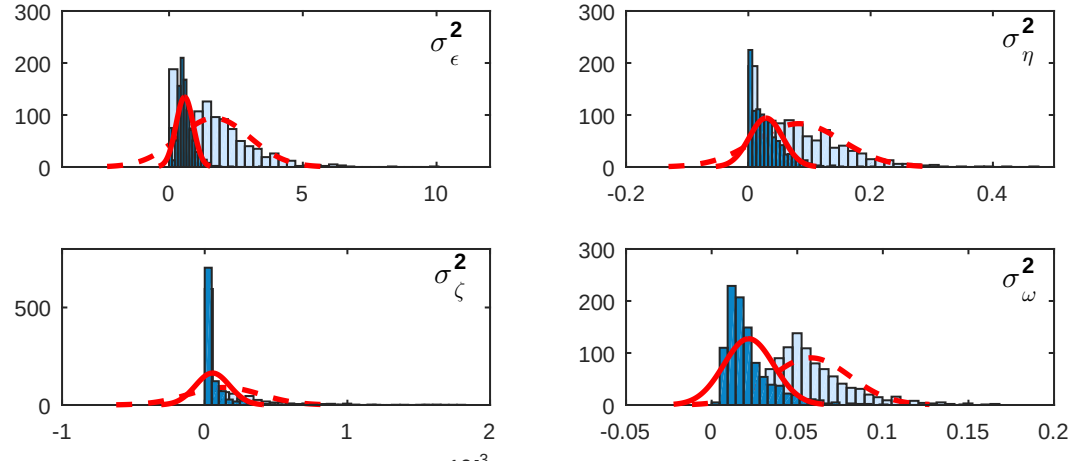
In the case of random AOs (Figure 3a), the distribution of the robust estimates of σ_ϵ^2 is concentrated around the true value 1. For σ_η^2, the mode of the distribution in the robust case occurs, in contrast to the ML case, at a value lower than the true one (0.08). This holds also for σ_ω^2, for which the mode underestimates the true value (0.05) in the



(a) AO



(b) AO patch



(c) IO

Figure 3: Distribution of the ML and robust estimates of the disturbance variances in the case of individual random AOs, a random patch of AOs and random IOs; ML estimates: light blue bars (histogram) and dashed red line (density), robust estimates: dark blue bars (histogram) and solid red line (density). The true parameter values correspond to the benchmark case: $\sigma^2_\epsilon = 1$, $\sigma^2_\eta = 0.08$, $\sigma^2_\zeta = 0.0001$, $\sigma^2_\omega = 0.05$.

robust case. However, it is also evident that the distribution of robust estimates of σ_η^2 is clearly less dispersed, which translates to a higher efficiency in the robust case, as could be seen in Table 1.

Similar observations as for AOs can be made also for random IOs (Figure 3c) with regards to all parameters and for a random AO patch (Figure 3b) except for σ_η^2. In the case of a random AO patch, ML leads to a substantial overestimation of σ_η^2. The distribution of the ML estimates of σ_η^2 is concentrated around the value 0.7, which is more than seven times higher than the true value (0.08). This large bias is due to the fact that outliers which occur in a group resemble a level shift, thereby corrupting the ML estimates of the level variance more than for individual outliers. The distribution of the robust estimates, in contrast, is more centered around the true value, which means that the robust method helps reduce the bias of the ML estimator.

To sum up, for all three considered outlier types the robust estimator delivers distribution of parameter estimates that is less dispersed than the distribution of ML estimates. In the benchmark scenario, which can be seen as a conservative case regarding the simulation results, the robust estimates of the irregular variance are unambiguously less biased than the ML estimates. For the variance of the seasonal disturbance, the ML estimator gives more accurate results, whereas for the variance of the level disturbance the results are more mixed.

Finally, we report on the performance of the robust AKF for detecting outliers. For this purposed, we compute the proportion of correctly adjusted outliers (also called sensitivity or true positive rate in the classification context) and focus thereby on individual AOs.[6] The results corresponding to our experiment design with randomly located and sized outliers are reported in Table 4. It can be observed that for the benchmark size of $\delta = 7 \cdot \text{PESD}$ the robust AKF is capable of correcting 95% to nearly 100% of outliers, depending on the variances scenario. If the reference outlier size is doubled, all outliers are identified and adjusted.

[6]We restrict the analysis to individual AOs for the following reasons. As regards AO patches, it is well–known that their effect on the series is very similar to that of a level shift lasting for a limited period. Detecting a level shift by capturing the single outlying observations in the time interval of the shift is very difficult; see, e.g., Marczak and Proietti (2016). Therefore, the proportion of identified AOs in a patch is expected to be lower than the proportion of identified single AOs. As regards IOs, differently than for the AO type, the proportion of the correctly adjusted outlier effects may not be an appropriate measure to evaluate an outlier detection method. While the effect of an AO lasts for one period only, the effect of an IO persists until the end of the series. Even though the effect of an IO at the time point of its occurrence would correspond to the effect of an AO, the effects at the subsequent time points may be very small and hardly detectable.

Table 4: Proportion of correctly adjusted outlier effects (in %) in the case of randomly located AOs. The outlier size at a location τ is given by $z_t\,\delta$.

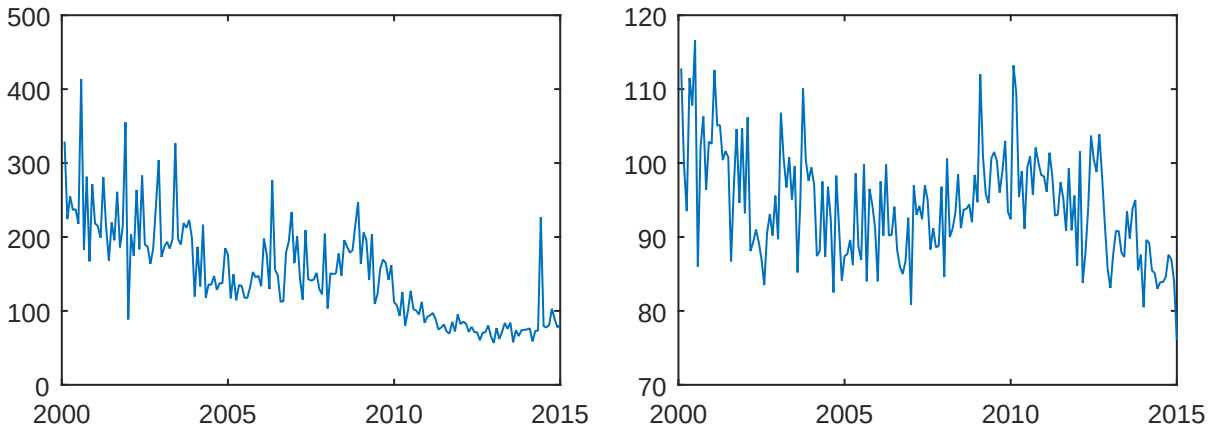
Size	Benchmark	sT–sS	uT–sS	sT–uS	uT–uS
$\delta = 7$·PESD	95.32	99.72	99.81	99.20	99.57
$\delta = 14$·PESD	100.00	100.00	100.00	100.00	100.00

6. Robust Forecasting: an Application

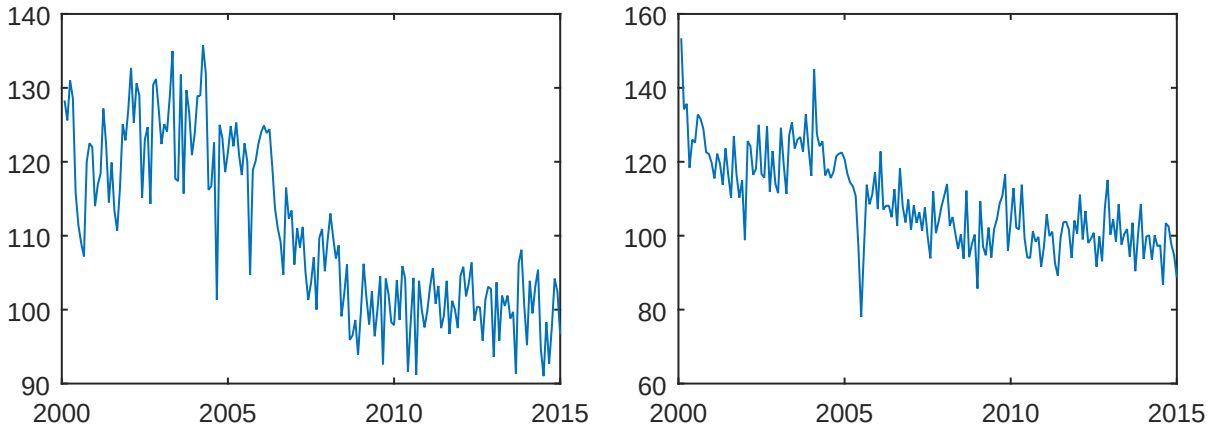
If the series is generated as $y_t = y_t^{\ddagger} + \xi_t$, with $y_t^{\ddagger}$ and ξ_t denoting uncontaminated latent series and outlier effects, respectively, robust forecasting deals with predicting $y_t^{\ddagger}$ using the observed past values of y_t. Two issues are involved: the first one is robust estimation of the model parameters; the second one deals with robustifying the forecasting method, so that a contaminated observation does not exert excessive influence on the forecast. Replacing y_t by an estimate of $y_t^{\ddagger}$ in the expression of the optimal linear predictor goes in this direction. The forecast application presented in this section aims at assessing the effectiveness of this strategy for forecasting a set of time series of trade flows. The perspective that is taken here is that we either believe that the outliers are not going to affect the future of y_t, or that we are inherently interested in forecasting the component $y_t^{\ddagger}$. If interest lies in predicting y_t, taking into account that outlier contamination can occur also in the future, so that the predictive density reflects the additional uncertainty due the presence of the outliers, other methods, such as modeling the distribution of the BSM disturbances as scale mixtures of normals, should be applied; see Bruce and Jurke (1996) and Bernardi et al. (2011).

6.1. Data

Our forecasting study focuses on a large set of monthly international trade data by BEC (Broad Economic Categories) product group classification for 28 members of the European Union (for a detailed information on the BEC classification, see: http://ec.europa.eu/eurostat/ramon/other_documents/bec/BEC_Rev_4.pdf.) From the original dataset, provided by Eurostat (download at: http://ec.europa.eu/eurostat/web/international-trade/data/database, label: `ext_st_28msbec`), we have extracted 540 time series available for the period 2000.M1 – 2014.M12, each having 180 observations. The series represent trade balances (indicated by the acronym `BAL_RT`), in the form of non–seasonally adjusted volume indices (`IVOL_NSA`), and are divided into two categories. The first one is the trade partner: euro area without Latvia and Lithuania (`EA17`), euro area without Lithuania (`EA18`), euro area with all current members (`EA19`), or European Union (`EU28`). The second category is related to the product type: capital goods (`CAP`), consumption goods



(a) Product type: `CAP`, Country: `FI` (Finland) **(b)** Product type: `CNS`, Country: `SE` (Sweden)



(c) Product type: `CTR`, Country: `HU` (Hungary) **(d)** Product type: `INT`, Country: `FR` (France)

Figure 4: Four series from the international trade statistics dataset for 28 members of the European Union. Each series depicted in the figure has the signature `BAL_RT,IVOL_NSA,EA17`, belongs to one of four product categories (`CAP`, `CNS`, `CTR`, `INT`), and represents one of 28 European Union countries.

(`CNS`), consumption goods plus motor spirit and passenger motor cars (`CTR`), intermediate goods (`INT`), or all products (`TOTAL`).

There are various reasons for considering international trade statistics for the purpose of this article. First, these series are typically contaminated by outliers which, at times, can be large. To illustrate the degree of contamination, four series taken from the dataset are plotted in Figure 4. Second, the dataset has large dimensions – a total of 45,361 series included in the original dataset. Outlier detection in such a case necessitates a procedure which allows for quick processing of each series. Although we focus on a subset of data for illustration purposes, our proposed method provides a simple and fast procedure for robust estimation and forecasting, and is capable of handling big data, like trade statistics. Finally, resorting to a large dataset enables a more reliable evaluation of the proposed data–cleaning procedure.

6.2. Design of the forecasting exercise and results

For all 540 series, we perform a pseudo real-time recursive forecasting exercise using the BSM with two specifications: in the first one (labelled "non–robust") no correction of the data takes place, whereas in the second one (labelled "robust") the robust AKF is applied for data cleaning. In the non-robust scenario, the exercise involves recursive estimation and prediction. In the robust case, apart from estimation and prediction, also outlier detection and correction is performed in a recursive manner. This means that data cleaning is not done beforehand, so that observations that are to be predicted (the test sample) are the same ones in both scenarios. This ensures the comparability of the results.

The training sample covers the time span 2000.M1 – 2009.M1. For a generic series, starting in 2009.M1 as the first forecast origin, we compute 1– to 12–period–ahead non–robust and robust forecasts. Then, for each next forecast origin until 2013.M12, the sample is extended by one month and 1– to 12–period–ahead forecasts are obtained for both specifications. This forecasting exercise yields for each forecast horizon from 1 to 12 a total of 540 forecasts for 60 time points from the respective time span. For example, 1–step–ahead forecasts are available in the time span 2009.M2 – 2014.M1, while the relevant time span for the 12–step–ahead forecast is 2010.M1 – 2014.M12.

In the evaluation of the forecasting precision in the non–robust and robust scenario, we focus on the assessment of the respective forecast densities instead of point forecasts. As shown by Ledolter (1989) in the ARIMA framework, point forecasts are largely unaffected by additive outliers, unless they occur close to the forecast origin. In contrast, outliers always inflate the variance of the prediction errors and hence the width of the prediction intervals. This means that even though the location of the forecast densities corresponding to contaminated and clean data may be very similar, the spread of the latter density is supposed to be smaller. The aim of the forecasting exercise is to assess whether data cleaning has a beneficial effect on the prediction uncertainty. The forecast densities are conditional on the recursive estimates of the parameters and assume Gaussianity.[7]

To compare the quality of densities associated with non–robust and robust forecasts, we apply scoring rules for evaluating the quality of density forecast. See Gneiting and Raftery (2007) for a comprehensive review of different scoring rules. In this article, we

[7]Giving up the Gaussianity assumption and accounting for parameter uncertainty would on the one hand provide more reliable forecast densities, on the other hand it would require computationally expensive algorithms, such as bootstrapping. In the case of a large set of series, the computational burden is too high relatively to the gain in terms of more reliable prediction densities. Moreover, our interest lies in the comparison of density forecasts in the non–robust and robust case, and as the densities are obtained based on the same assumptions, the uncertainty will be underestimated in both cases.

adopt two proper scores: the log score, proposed by Good (1952), and the continuous ranked probability score (CRPS), introduced by Matheson and Winkler (1976).[8] The log score (LogS) at the realized outcome y_t is given as:

$$\text{LogS}(y_t) = \log f(y_t),$$

where $f(u)$ denotes a density forecast with the corresponding cumulative distribution function $F(u)$. Despite the desirable properties of the log score and its widespread use in the evaluation of density forecasts, one drawback is its lack of robustness. For example, in the presence of outliers, if for a single observation the forecast density is completely missing the realized outcome, the log score attached to this observation approaches $-\infty$. As a consequence, if a forecasting model is evaluated based on an average score, then the log scoring rule discredits this model, even if its overall forecasting performance at all other observations might be good. Therefore, to safeguard against the possible non–robustness of the log score, we also adopt the CRPS—a more robust and tolerant scoring rule which assigns high numerical scores for probabilities at values close, but not necessarily equal, to the realized one. CRPS penalizes deviations of the predictive cumulative distribution function from the true one for a particular time point. More formally,

$$\text{CRPS}(y_t) = -\int [F(u) - I(y_t)]^2 \, \mathrm{d}u, \tag{24}$$

where $I(\cdot)$ is an indicator variable taking value 1, if $u > y_t$, and 0 otherwise. Based on the assumption of Gaussian predictive distribution, we can use an analytical formula for the computation of the CRPS; see Gneiting and Raftery (2007)[9]:

$$\text{CRPS}(y_t) = \sqrt{\tilde{f}_t}\left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{y_t - \tilde{y}_{t|t-h}}{\sqrt{\tilde{f}_t}}\right) - \frac{y_t - \tilde{y}_{t|t-h}}{\sqrt{\tilde{f}_t}}\left(2\Phi\left(\frac{y_t - \tilde{y}_{t|t-h}}{\sqrt{\tilde{f}_t}}\right) - 1\right)\right],$$

where $\tilde{y}_{t|t-h}$ denotes the prediction of y_t based on the information set in period $t-h$ (with $h = 1, \ldots, 12$ in our forecasting experiment), and $\sqrt{\tilde{f}_t}$ is the root mean square error (RMSE) of $\tilde{y}_{t|t-h}$.

To facilitate the comparison between the non–robust and robust forecasts, we consider

[8] A scoring rule is proper if the expected value of the score is maximized when the forecast distribution is the same as the data generating distribution.

[9] Please note that otherwise one could resort to the formulation $\text{CRPS}(y_t) = \frac{1}{2}\,\text{E}_F|Y - Y'| - \text{E}_F|Y - y_t|$, where E_F is the expectation with respect to the forecast distribution F, and Y and Y' are independent random draws from F. Then, CRPS can be computed using algorithms for approximating the previous formula; see Panagiotelis and Smith (2008).

the differences in the score values (log score or CRPS), i.e. we subtract the score value in the non–robust case from the corresponding score value in the robust case. Both the log score and the CRPS are defined in a way that higher values (less negative in the CRPS case) indicate a lower dispersion and thus better predictive performance. Therefore, positive score differences provide evidence for the superiority of the data cleaning approach. The forecasting experiment based on the full set of series results, for a particular forecast horizon from 1 to 12, in the distribution of the log score and CRPS differences given at each time point of the evaluation sample. The empirical distribution of the score differences based on 540 series are smoothed using a Gaussian kernel.

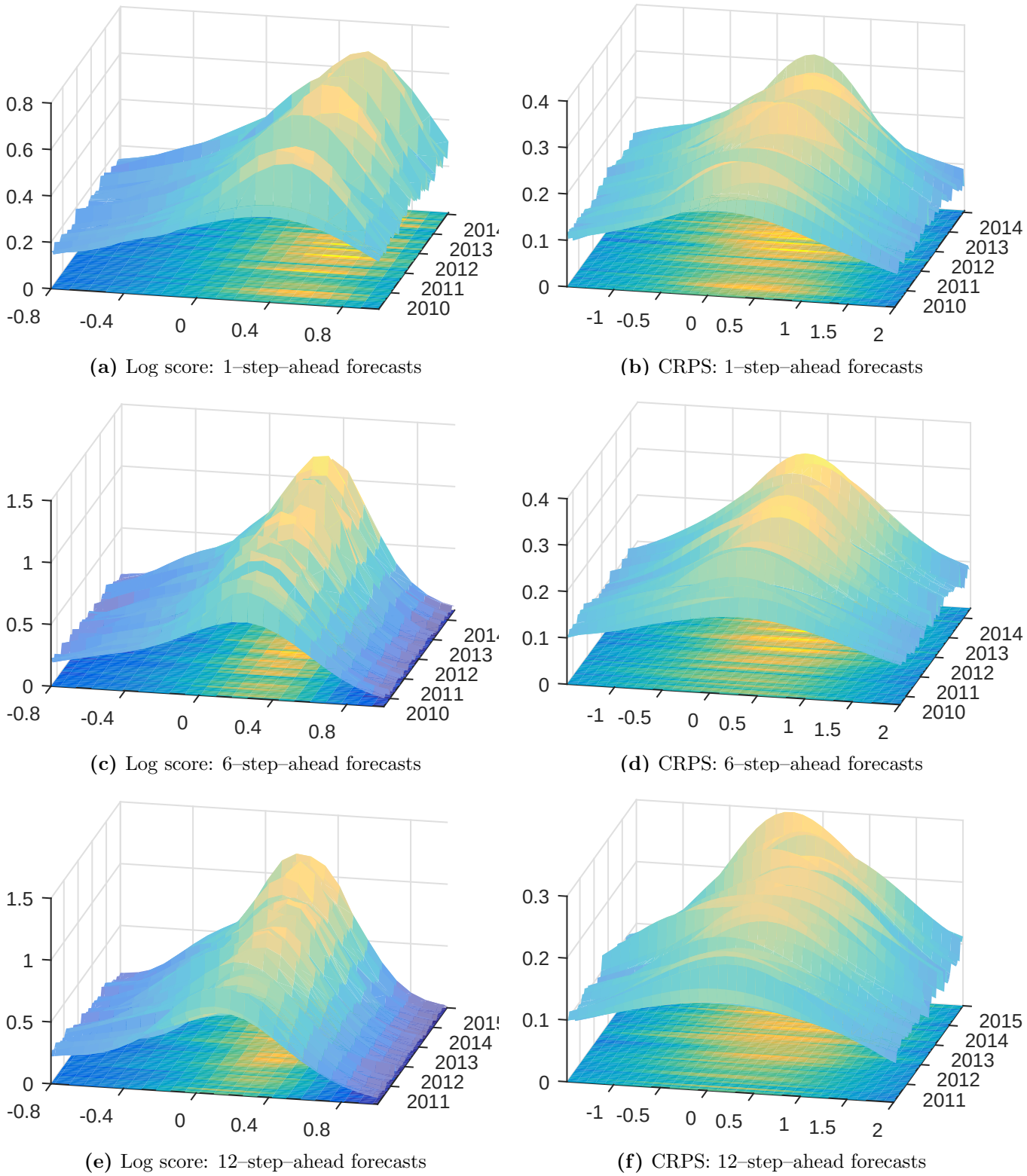
The smoothed distributions of differences between scores corresponding to the robust and non–robust forecasts are depicted in Figure 5. The selected forecast horizons are 1, 6 and 12 months. The height of the surface plot at each point as well as the color along with the hue give the information about the probability value. The color scale ranges from dark indicating the lowest probabilities to the yellow indicating the highest probabilities. To make the results easier to interpret, each three–dimensional surface plot is projected on a two–dimensional plane.

It is evident that for both scoring rules and all forecast horizons, the mass of the distribution at any of the time points is concentrated at positive values of the score differences. This means that the scores associated with the robust forecasts are higher or, in other words, forecasts obtained with cleaned data are considered superior. The results in favor of the robust alternative are, independent of the forecast horizon, more pronounced in the CRPS case. As regards the log score, the superiority of the robust alternative is especially visible in the case of 1–period–ahead forecasts. All these findings confirm that the robust AKF is a procedure which for the considered set od international trade statistics effectively cleans the data, which is reflected in the reduced forecast uncertainty and in better calibrated predictive densities.

7. Conclusions

This paper has developed a robust augmented Kalman filter (AKF) for data cleaning, enabling robust estimation of the parameters of a possibly non-stationary state space model featuring fixed effects. The main idea is based on shrinking the updated, or real time, estimates towards the one–step–ahead predictions, in the presence of an outlier. Our methodology combines the approach of the robust Kalman filter by Masreliez and Martin (1977) with the augmentation approach for the Kalman filter accounting for the presence of nonstationary elements and regression effects (de Jong, 1991).

For the purpose of evaluating the proposed method we have focused on the class of



(a) Log score: 1–step–ahead forecasts

(b) CRPS: 1–step–ahead forecasts

(c) Log score: 6–step–ahead forecasts

(d) CRPS: 6–step–ahead forecasts

(e) Log score: 12–step–ahead forecasts

(f) CRPS: 12–step–ahead forecasts

Figure 5: Distribution of differences in the scores based on the robust and non–robust forecasts for 540 European trade statistics series. The robust (non–robust) forecasts are obtained in a pseudo real-time recursive forecasting exercise using cleaned (original) data. Data cleaning is performed recursively for each forecast origin. Training sample: 2000.M1 – 2009.M1; evaluation sample: 2009.2 – 2014.M12. The rows indicate the corresponding forecast horizons: (from top to bottom) 1–, 6,–, and 12–step–ahead forecasts. Left panel: log score differences; right panel: CRPS differences. The kernel density estimates are obtained by using a Gaussian kernel.

structural time series models and, in particular, on the basic structural model (BSM) which is a popular model used for the analysis of seasonal time series. In the applicative part of the paper, we have conducted a Monte Carlo experiment, in which the series, generated according to the BSM with different parameter settings, are affected by random additive outliers (AOs) occurring either individually or in a patch, and random innovation outliers (IOs). The results show that for all three types of outlier contamination the robust estimates of the parameters are, in general, more accurate.

Secondly, we have assessed the gains arising from the robustification for out-of-sample prediction, by conducting a recursive forecasting exercise on a large and representative set of foreing trade time series, characterized by high levels of outlier contamination. We have compared the density forecasts with and without the data cleaning using two scoring rules: the log score and the continuous ranked probability score; we have concluded that data cleaning with the robust AKF produces more accurate and better calibrated density forecasts. Our overall conclusion is that the robust AKF is a particularly attractive tool for handling large sets of series, like the international trade statistics database.

References

References

Bernardi, M., Della Corte, G., Proietti, T., 2011. Extracting the cyclical component in hours worked. Studies in Nonlinear Dynamics & Econometrics 15 (3).

Bianco, A. M., Garcia Ben, M., Martinez, E., Yohai, V. J., 2001. Outlier Detection in Regression Models with ARIMA Errors Using Robust Estimates. Journal of Forecasting 20 (8), 565–579.

Bruce, A. G., Jurke, S. R., 1996. Non-gaussian seasonal adjustment: X-12-arima versus robust structural models. Journal of Forecasting 15 (4), 305–327.

de Jong, P., 1991. The Diffuse Kalman Filter. The Annals of Statistics 19, 1073–1083.

Francke, M. K., Koopman, S. J., De Vos, A. F., 2010. Likelihood Functions for State Space Models with Diffuse Initial Conditions. Journal of Time Series Analysis 31 (6), 407–414.

Gandhi, M. A., Mili, L., 2010. Robust Kalman Filter Based on a Generalized Maximum–Likelihood–Type Estimator. IEEE Transactions on Signal Processing 58 (5), 2509–2520.

Gneiting, T., Raftery, A. E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association 102 (477), 359–378.

Good, I. J., 1952. Rational Decisions. Journal of the Royal Statistical Society: Series B 14 (1), 1107–114.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York.

Harvey, A. C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.

Harvey, A. C., 2013. Dynamic Models for Volatility and Heavy Tails: with Applications to Financial and Economic Time Series. No. 52. Cambridge University Press.

Harvey, A. C., Todd, P. H. J., 1983. Forecasting Econometric Time Series with Structural and Box-Jenkins Models (with discussion). Journal of Business and Economic Statistics 1 (4), 299–315.

Henderson, H. V., Searle, S. R., 1981. On Deriving the Inverse of a Sum of Matrices. Siam Review 23 (1), 53–60.

Ledolter, J., 1989. The Effect of Additive Outliers on the Forecasts from ARIMA Models. International Journal of Forecasting 5, 231–240.

Liu, H., Shah, S., Jiang, W., 2004. On–line Outlier Detection and Data Cleaning. Computers and Chemical Engineering 28, 1635–1647.

Marczak, M., Proietti, T., 2016. Outlier Detection in Structural Time Series Models: The Indicator Saturation Approach. International Journal of Forecasting 32 (1), 180–202.

Maronna, R., Martin, D., Yohai, V., 2006. Robust Statistics: Theory and Methods. John Wiley & Sons, Chichester. ISBN.

Martin, R. D., 1979. Robust Estimation for Time Series Autoregressions. In: Launer, R. L., Wilkinson, G. N. (Eds.), Robustness in Statistics. Academic Press, New York, pp. 147–176.

Martin, R. D., Thomson, D. J., 1982. Robust–resistant Spectrum Estimation. Proceedings of the IEEE 70, 1097–1115.

Masreliez, C. J., Martin, R. D., 1977. Robust Bayesian Estimation for the Linear Model and Robustifying the Kalman Filter. IEEE Transactions on Automatic Control AC–22, 361–371.

Matheson, J. E., Winkler, R. L., 1976. Scoring Rules for Continuous Probability Distributions. Management Science 22 (10), 1087–1096.

Panagiotelis, A., Smith, M., 2008. Bayesian Density Forecasting of Intraday Electricity Prices Using Multivariate Skew t Distributions. International Journal of Forecasting 24, 710–727.

Rosenberg, B., 1973. Random Coefficients Models: the Analysis of a Cross Section of Time Series by Stochastically Convergent Parameter Regression. In: Annals of Economic and Social Measurement. Vol. 2. NBER, pp. 399–428.

Ruckdeschel, P., Spangl, B., Pupashenko, D., 2014. Robust Kalman Tracking and Smoothing with Propagating and Non–propagating Outlier. Statistical Papers 55, 93–123.