KU LEUVEN
FACULTY OF ECONOMICS AND BUSINESS

Robust estimation of linear state space models

Crevits R, Croux C.

KBI_1713

Robust Estimation of Linear State Space Models

Ruben Crevits and Christophe Croux

Faculty of Economics and Business, KU Leuven

Abstract. The model parameters of linear state space models are typically estimated with maximum likelihood estimation, where the likelihood is computed analytically with the Kalman filter. Outliers can deteriorate the estimation. Therefore we propose an alternative estimation method. The Kalman filter is replaced by a robust version and the maximum likelihood estimator is robustified as well. The performance of the robust estimator is investigated in a simulation study. Robust estimation of time varying parameter regression models is considered as a special case. Finally, the methodology is applied to real data.

Keywords. Kalman Filter, Forecasting, Outliers, Time varying parameters.

1 Introduction

Linear state space models are used for a wide range of applications. The most convenient way to estimate the parameters of such a model is by computing the likelihood with the Kalman filter and then maximize this likelihood, assuming normality of the noise. The maximum likelihood estimator is very common for the estimation of such models (Brockwell and Davis, 2002; Durbin and Koopman, 2012). There also exist Bayesian methods to estimate state space models, but these are not considered here.

This paper proposes a way to estimate the model parameters (also called hyperparameters or static parameters) of state space models robustly. The Kalman filter needed to compute the likelihood cannot cope with outliers, and therefore needs to be replaced by a robust filter. Several proposals for robust filters have been made in the literature, and we choose the robust filter of Cipra (1997).

Robust estimation of model parameters is less studied. In many applications the model parameters are supposed to be known, but in practice this is often not the case. Only few robust estimation procedures have been proposed. Agamennoni et al. (2011) do maximum likelihood estimation assuming multivariate t-distributed noise. Harvey and Luati (2014) developed a dynamic conditional score model based on the t-distribution. Their method only works for univariate time series. We should mention that Muler et al. (2009) developed a robust estimator of the general ARMA model, which is related to our approach for estimating linear state space models.

The structure of the paper is as follows. In Section 2 the linear state space model with its Kalman filter and maximum likelihood estimator are introduced. In Section 3 we propose a robust approach. The different estimators are compared in a simulation study in Section 4. The special case of the time varying parameter models is considered in Section 5: the performance of the robust estimator is studied for simulated and real data sets.

2 Linear Gaussian state space models

In a state space model we assume that a time series $\mathbf{y}_t$ is generated from a series of unobserved states $\boldsymbol{\theta}_t$. The distribution of the states follows from specifying the initial state density and the transition density. In a linear Gaussian state space model all distributions are normal and the expected value of the state only depends linearly on the previous state. The observations are generated from the states through the observation density. The general linear Gaussian state space model for a multivariate time series $\mathbf{y}_t$ is described by:

$$\begin{array}{lrcll} \text{initial state density} & p(\boldsymbol{\theta}_0) & \sim & \mathcal{N}(\boldsymbol{\theta}_{0|0}, \mathbf{P}_{0|0}) & \\ \text{transition density} & p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}) & \sim & \mathcal{N}(\mathbf{F}_t\boldsymbol{\theta}_{t-1}, \boldsymbol{\Lambda}_t) & \text{for } t \geq 1 \\ \text{observation density} & p(\mathbf{y}_t|\boldsymbol{\theta}_t) & \sim & \mathcal{N}(\mathbf{H}_t\boldsymbol{\theta}_t, \boldsymbol{\Sigma}_t) & \text{for } t \geq 1. \end{array} \tag{2.1}$$

The parameters are $\boldsymbol{\theta}_{0|0}$, $\mathbf{P}_{0|0}$, $\mathbf{F}_t$, $\boldsymbol{\Lambda}_t$, $\mathbf{H}_t$ and $\boldsymbol{\Sigma}_t$. Often they are taken to be constant over time. We assume that they can be written as a function of an unknown multidimensional parameter $\boldsymbol{\phi}$. A common notation for the same model is:

$$\begin{aligned} \boldsymbol{\theta}_t &= \mathbf{F}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t \\ \mathbf{y}_t &= \mathbf{H}_t\boldsymbol{\theta}_t + \mathbf{v}_t \end{aligned}$$

where the innovation noise $\mathbf{w}_t$ follows a $\mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}_t)$ and the observation or measurement noise $\mathbf{v}_t$ is distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)$.

An example is the time invariant univariate linear state space model with $\boldsymbol{\theta}_{0|0} = 0$ and $\mathbf{P}_{0|0} = 10^2$. This is called a diffuse initial state density (Durbin and Koopman, 2012). By setting $\mathbf{H}_t = 1$, the states are the expected values of the observations. With $\mathbf{F}_t = F$, $\boldsymbol{\Lambda}_t = \lambda^2$ and $\boldsymbol{\Sigma}_t = \sigma^2$, the vector of unknown parameters is chosen as $\boldsymbol{\phi} = (\log\sigma, \log\lambda, F)'$, and varies in a 3-dimensional space.

Denote the observed time series of dimension d as

$$\mathbf{y}_{1:T} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}.$$

The unknown model parameter $\boldsymbol{\phi}$ can be estimated using the maximum likelihood estimator:

$$\hat{\boldsymbol{\phi}}_{\text{MLE}} = \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \log p(\mathbf{y}_{1:T}|\boldsymbol{\phi}) = \underset{\boldsymbol{\phi}}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\phi})$$

with $p(\mathbf{y}_1|\mathbf{y}_{1:0},\boldsymbol{\phi}) := p(\mathbf{y}_1|\boldsymbol{\phi})$. In above expression, the predictive density still has to be evaluated; it is normal with mean $\hat{\mathbf{y}}_{t|t-1}$ and covariance $\mathbf{S}_t$. The Kalman filter allows to compute this mean and covariance sequentially. The Kalman recursions are derived analytically, e.g. Petris et al. (2009), and given by

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{t|t-1} &= \mathbf{F}_t\hat{\boldsymbol{\theta}}_{t-1|t-1} \\
\hat{\mathbf{y}}_{t|t-1} &= \mathbf{H}_t\hat{\boldsymbol{\theta}}_{t|t-1} \\
\mathbf{P}_{t|t-1} &= \mathbf{F}_t'\mathbf{P}_{t-1|t-1}\mathbf{F}_t + \boldsymbol{\Lambda}_t \\
\mathbf{S}_t &= \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \boldsymbol{\Sigma}_t \\
\mathbf{K}_t &= \mathbf{P}_{t|t-1}\mathbf{H}_t'\mathbf{S}_t^{-1} \\
\hat{\boldsymbol{\theta}}_{t|t} &= \hat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right) \\
\mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}_t'\mathbf{S}_t^{-1}\mathbf{H}_t\mathbf{P}_{t|t-1}.
\end{aligned} \tag{2.2}$$

The likelihood $p(\mathbf{y}_t|\mathbf{y}_{1:t-1},\boldsymbol{\phi})$ is found by evaluating the density function of a normal with mean $\hat{\mathbf{y}}_{t|t-1}$ and covariance $\mathbf{S}_t$ at $\mathbf{y}_t$. Note that all expressions in (2.2) are conditional on the model parameter $\boldsymbol{\phi}$. The negative log-likelihood, apart from a constant term, is

$$\frac{1}{2T}\sum_{t=1}^{T}\Big(\log(\det\mathbf{S}_t(\boldsymbol{\phi})) + \left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})\right)'\mathbf{S}_t(\boldsymbol{\phi})^{-1}\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})\right)\Big). \tag{2.3}$$

The likelihood depends on $\boldsymbol{\phi}$ through $\hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})$ and $\mathbf{S}_t(\boldsymbol{\phi})$. The objective function has to be optimized numerically. We use the Nelder-Mead algorithm implemented in the function `optim` in `R`. A good initial value of $\boldsymbol{\phi}$ is needed to start up the numerical optimizer, as the optimization problem is not convex in general.

The MLE is not robust against observation outliers in two ways. Suppose $\mathbf{y}_t$ is an outlier. (i) The second term in the log-likelihood (2.3) becomes large. The estimator $\hat{\boldsymbol{\phi}}_{\text{MLE}}$ is possibly seriously affected by this outlier. (ii) The Kalman recursions yield that the prediction of the next observation is

$$\hat{\mathbf{y}}_{t+1|t} = \mathbf{H}_{t+1}\mathbf{F}_{t+1}\Big(\hat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}\right)\Big).$$

One sees that an outlier $\mathbf{y}_t$ has a large influence on the prediction of the next value. Therefore, we need robust filter recursions.

3 A robust approach

In the first subsection we review a suitable robust Kalman filter. In the second subsection we propose two ways to robustify the maximum likelihood estimator.

3.1 Robust filtering

The robustified Kalman filter of Cipra (1997) is inspired by an alternative derivation of the Kalman recursions. The state prediction $\hat{\boldsymbol{\theta}}_{t|t}$ is equal to the solution of a least squares problem, as shown in Appendix A:

$$\hat{\boldsymbol{\theta}}_{t|t} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \left(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}\right)' \mathbf{P}_{t|t-1}^{-1} \left(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}\right) + \left(\mathbf{y}_t - \mathbf{H}_t\boldsymbol{\theta}\right)' \boldsymbol{\Sigma}_t^{-1} \left(\mathbf{y}_t - \mathbf{H}_t\boldsymbol{\theta}\right) \right\}, \tag{3.1}$$

which is equivalent to

$$\hat{\boldsymbol{\theta}}_{t|t} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \left(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}\right)' \mathbf{P}_{t|t-1}^{-1} \left(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}\right) + \sum_{i=1}^{d} \left(s_{it} - \mathbf{b}_{it}\boldsymbol{\theta}\right)^2 \right\} \tag{3.2}$$

with $\mathbf{s}_t = \boldsymbol{\Sigma}_t^{-1/2}\mathbf{Y}_t$ and $\mathbf{b}_t = \boldsymbol{\Sigma}_t^{-1/2}\mathbf{H}_t$; $\mathbf{b}_{it}$ is the i-th row of the matrix $\mathbf{b}_t$. The squares in (3.2) are replaced by another loss function:

$$\hat{\boldsymbol{\theta}}_{t|t} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left\{ \frac{1}{2} \left(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}\right)' \mathbf{P}_{t|t-1}^{-1} \left(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}\right) + \sum_{i=1}^{d} \rho\left(s_{it} - \mathbf{b}_{it}\boldsymbol{\theta}\right) \right\} \tag{3.3}$$

with ρ a loss function that is less sensitive to outliers. Cipra (1997) shows that an approximate solution of (3.3) leads to the Kalman recursions, but with

$$\mathbf{S}_t = \mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \boldsymbol{\Sigma}_t^{1/2}\mathbf{W}_t^{-1}\boldsymbol{\Sigma}_t^{1/2}, \tag{3.4}$$

where $\mathbf{W}_t = \operatorname{diag}(w_{1t}, w_{2t}, \ldots, w_{dt})$ is a diagonal matrix of weights:

$$w_{it} = \frac{\psi\left(s_{it} - \mathbf{b}_{it}\hat{\boldsymbol{\theta}}_{t|t-1}\right)}{s_{it} - \mathbf{b}_{it}\hat{\boldsymbol{\theta}}_{t|t-1}}, \tag{3.5}$$

with ψ be the derivative of ρ. We use the Huber ψ-function:

$$\psi_{\mathrm{H}}(\mathbf{x}) = (\psi_{\mathrm{H}}(x_1), \psi_{\mathrm{H}}(x_2),\ \ldots\ )', \qquad \psi_{\mathrm{H}}(x_i) \begin{cases} x_i & \text{if } |x_i| < k \\ \operatorname{sign}(x_i)k & \text{otherwise} \end{cases}$$

with $k = 2$. This adaptation makes the variance (3.4) larger if there is an outlier at time t.

3.2 Robust estimation of model parameters

We present two robust procedures to estimate the model parameters: the Huber maximum likelihood and the maximum trimmed likelihood.

Huber maximum likelihood

The Huber maximum likelihood estimator is

$$\hat{\boldsymbol{\phi}}_{\mathrm{H}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \frac{1}{2T} \sum_{t=1}^{T} \log(\det \mathbf{S}_t(\boldsymbol{\phi})) + \frac{c_H}{T} \sum_{t=1}^{T} \rho_H \left(\mathbf{S}_t(\boldsymbol{\phi})^{-1/2} \left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi}) \right) \right), \tag{3.6}$$

where $\hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})$ and $S_t(\boldsymbol{\phi})$ are computed with the robust filter of Cipra (1997) outlined in the previous section. The quadratic function in the likelihood of (2.3) is replaced by the multivariate Huber ρ-function (Hampel et al., 1986):

$$\rho_H(\mathbf{x}) = \begin{cases} \frac{1}{2}||\mathbf{x}||^2 & \text{if } ||\mathbf{x}|| < k \\ k||\mathbf{x}|| - \frac{k^2}{2} & \text{otherwise.} \end{cases} \tag{3.7}$$

If $\boldsymbol{\phi}$ is the true parameter, $\mathbf{x} = \mathbf{S}_t(\boldsymbol{\phi})^{-1/2} \left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi}) \right)$ is multivariate standard normal, and $||\mathbf{x}||$ follows a χ_d distribution with d degrees of freedom. Therefore we choose $k = F_{\chi_d}^{-1}(0.95)$, which is about 2 for univariate observations. The constant c_H is such that expected value of the objective function in (3.6) is the same as in (2.3). For its computation we refer to Appendix B.

The Huber ρ-function is quadratic for values of $||\mathbf{x}||$ smaller than k, but depends only linearly on $||\mathbf{x}||$ for values larger than k. Consequently this loss function is less influenced by large errors than the usual quadratic loss function. Notice that a bounded ρ-function cannot be taken, since the optimization problem in (3.6) would have a degenerate solution where $\mathbf{S}_t(\boldsymbol{\phi})$ is zero. If $\rho_H = \frac{1}{2}||\mathbf{x}||^2$, and the Kalman filter is used, then the estimator is equal to the maximum likelihood estimator of the model in (2.1).

Maximum trimmed likelihood

For all values of $\mathbf{y}_t$ the Mahalanobis distance to the prediction $\hat{\mathbf{y}}_{t|t-1}$ is computed:

$$d_t(\boldsymbol{\phi}) = \left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})\right)' \mathbf{S}_t(\boldsymbol{\phi})^{-1} \left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})\right).$$

The fraction α (f.e. $\alpha = 0.1$) of the observations with the highest Mahalanobis distances d_t are not considered in the maximum likelihood estimation. The estimator is

$$\hat{\boldsymbol{\phi}}_T = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \frac{1}{2T(1-\alpha)} \sum_{t\in\mathcal{D}} \Big(\log(\det \mathbf{S}_t(\boldsymbol{\phi})) + c_T \left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})\right)' \mathbf{S}_t(\boldsymbol{\phi})^{-1} \left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})\right)\Big). \tag{3.8}$$

with $\mathcal{D}$ the set of epochs t with d_t smaller than the $(1-\alpha)$ largest distance d_t. The values $\hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})$ and $S_t(\boldsymbol{\phi})$ are computed with the robust filter of Cipra (1997) from Subsection 3.1. The constant c_T makes the expected value of the trimmed log-likelihood equal to that of the untrimmed log-likelihood at the true model parameter. Its computation is given in Appendix B.

4 Simulations

The estimators are compared in different settings: a univariate model in the first subsection and a multivariate model in the second subsection.

4.1 Univariate state space model

Consider the following model with univariate states and observations:

$$\begin{aligned} \theta_t &= F\theta_{t-1} + w_t \\ y_t &= \theta_t + v_t. \end{aligned} \tag{4.1}$$

with $w_t \sim \mathcal{N}(0, \lambda^2)$, $v_t \sim \mathcal{N}(0, \sigma^2)$, $F = 1$, $\sigma = 1$, $\lambda = 0.1$, $\theta_0 = 0$. There is no coefficient before θ_t in the observation equation in order to keep the model parameters identifiable. We use an estimation period of length $T = 100$ and a subsequent test period of $T_o = 100$. The outliers are generated by replacing σ, the standard deviation of the observation noise, by a random variable that has 10% chance of being 10σ and 90% chance of being σ. These outliers are only added in the estimation period. We do $N = 1000$ simulations.

The parameter vector is $\boldsymbol{\phi} = (\log \sigma, \log \lambda, F)$ and we choose a diffuse initial state density with $\theta_{0|0} = 0$ and $P_{0|0} = 10^2$. The logarithmic transform is done to avoid setting boundaries on the parameter space. An estimate $\hat{\boldsymbol{\phi}}$ is computed from the estimation period[1].

We compare the one-step ahead prediction errors in the out-of-sample period for the three estimators. The out-of-sample mean squared error (MSE) is

$$\text{MSE} = \frac{1}{T_o} \sum_{t=T+1}^{T+T_o} (y_t - \hat{y}_{t|t-1}(\hat{\boldsymbol{\phi}}))^2. \tag{4.2}$$

Table 1 reports this MSE, averaged over 1000 simulations, for the MLE and the robust alternatives based on the Huber loss function (3.6) and based on trimming (3.8). The maximum likelihood estimator is the best for clean data, where no outliers are present. However, the robust estimators perform barely worse. With the contaminated time series the maximum trimmed likelihood estimator is the best.

Table 1: Out-of-sample MSE for a univariate state space model, averaged over 1000 simulations.

	MLE	Huber	Trimmed
clean	1.73	1.73	1.82
contaminated	5.08	2.47	2.08

4.2 Multivariate state space model

We do a similar exercise for a multivariate state space model. Consider the following model

$$\begin{aligned} \boldsymbol{\theta}_t &= \boldsymbol{\theta}_{t-1} + \mathbf{w}_t \\ \mathbf{y}_t &= \boldsymbol{\theta}_t + \mathbf{v}_t \end{aligned}$$

[1] The initial value for the numerical optimizer is set at the true value of $\boldsymbol{\phi}$, for all simulations to come.

where $\mathbf{v}_t$ and $\mathbf{w}_t$ are normal with zero mean and variance $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda} = q\boldsymbol{\Sigma}$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{11} & \rho\sqrt{\Sigma_{11}\Sigma_{22}} \\ \rho\sqrt{\Sigma_{11}\Sigma_{22}} & \Sigma_{22} \end{bmatrix}.$$

The unknown parameters are Σ_{11}, Σ_{22}, ρ and q. The variances Σ_{11} and Σ_{22} should be positive and the correlation ρ should be in [-1,1]. We restrict q to be in $[0, 1]$. The parameter vector is a transformation of these parameters such that the restrictions are automatically satisfied:

$$\boldsymbol{\phi} = \left(\frac{1}{2}\log(\Sigma_{11}), \frac{1}{2}\log(\Sigma_{22}), \log\left(\frac{1+\rho}{1-\rho}\right), \log\left(\frac{q}{1-q}\right)\right)'.$$

We generate the data from a model with $\Sigma_{11} = \Sigma_{22} = 0.25$, $\rho = 0.5$ and $q = 0.01$. We take an estimation period of $T = 200$ and an out-of-sample period of $T_o = 100$. We generate outliers in the estimation period as in the previous section: by replacing $\boldsymbol{\Sigma}$ by a random variable that has 10% chance of being

$$\boldsymbol{\Sigma}_{\text{out}} = 100 \begin{bmatrix} 25 & -24 \\ -24 & 25 \end{bmatrix}$$

and 90% chance of being $\boldsymbol{\Sigma}$. We compute

$$\text{MSE} = \frac{1}{T_o}\sum_{t=T+1}^{T+T_o} (\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\hat{\boldsymbol{\phi}}))'(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\hat{\boldsymbol{\phi}})) \tag{4.3}$$

for each simulated series. The averages of the MSE over $N = 1000$ simulations are reported in Table 2.

For clean data, the robust estimators have almost the same out-of-sample MSE as the maximum likelihood estimator. The difference is negligible. For contaminated data, the robust estimators outperform the MLE, and the trimmed maximum likelihood estimator has the best MSE. The MLE is heavily affected by the outliers in the estimation period, and that is reflected in the performance in the out-of-sample period.

5 Time varying parameter models

The linear regression model is widely used, but has its limitations: the constant coefficient may in fact be time varying. Stock and Watson (1996) show that the regression model with

Table 2: Out-of-sample MSE for a multivariate state space model, averaged over 1000 simulations.

	MLE	Huber	Trimmed
clean	0.282	0.283	0.283
contaminated	0.989	0.342	0.314

time varying coefficients has a good performance for forecasting economic time series. In this section the time varying regression model is estimated in a robust way.

Consider the model

$$y_t = c + \mathbf{x}_t' \boldsymbol{\theta}_t + v_t, \tag{5.1}$$

where y_t is the response variable, $\boldsymbol{x}_t$ the vector of predictor variables, c the intercept and v_t a normally distributed error term with mean 0 and variance σ^2, for $1 \leq t \leq T$. If the slope coefficient $\boldsymbol{\theta}_t$ is time invariant, the model reduces to a linear regression model, and is typically estimated by the Ordinary Least Squares (OLS) estimator. The latter estimator is not robust.

A common way to allow for time varying parameters is to model them as a random walk, resulting in stochastic regression coefficients. This was already done in the early seventies by Rosenberg (1972). We consider a linear regression model with time varying coefficients changing over time like a random walk:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \mathbf{w}_t. \tag{5.2}$$

The innovations $\mathbf{w}_t$ have normal distributions with mean zero and a covariance matrix $\boldsymbol{\Lambda}$ with usually only diagonal nonzero elements. Note that (5.1) and (5.2) are a special case of the linear state space model of (2.1). This can be easily seen by taking $\mathbf{F}_t$ the identity matrix, and $\mathbf{H}_t = \mathbf{x}_t'$.

Table 3: Contamination schemes. The outlier probability ε is 0.10.

	Description
Clean data	$v_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
Symmetric outliers	$v_t \overset{\text{iid}}{\sim} (1-\varepsilon)\mathcal{N}(0, \sigma^2) + \varepsilon\mathcal{N}(0, 100\sigma^2)$.
Asymmetric outliers	$v_t \overset{\text{iid}}{\sim} (1-\varepsilon)\mathcal{N}(0, \sigma^2) + \varepsilon\mathcal{N}(10\sigma, \sigma^2)$.
Bad leverage points	with probability $(1-\varepsilon)$: $x_t \sim \mathcal{N}(2, 1)$ and $v_t \sim \mathcal{N}(0, \sigma^2)$
	with probability ε: $x_t \sim \mathcal{N}(2, 100)$ and $v_t \sim \mathcal{N}(0, 100\sigma^2)$.

5.1 Simulation results

We apply the time varying model for simulated time series. We take a univariate x_t. The parameters are collected in $\boldsymbol{\phi} = (c, \log\sigma, \log\lambda)'$ with $\boldsymbol{\Lambda} = \lambda^2$. The mean and variance of the initial state density is chosen 0 and 10^6, respectively, as in Stock and Watson (1996). This is again the diffuse prior from Durbin and Koopman (2012).

We simulate 1000 time series generated by the model defined in equations (5.1) and (5.2) starting with $\theta_0 = 0$. We take as intercept $c = 1$ and as variance parameters are $\sigma = 1$ and $\lambda = 0.01$. We choose the covariate $x_t \sim \mathcal{N}(2, 1)$. The parameters are chosen to have a high enough signal to noise ratio. The variance λ^2 is sufficiently high to have a detectable time varying trend. The length of the estimation period is $T = 100$. The out-of-sample period ranges from $T + 1$ till $T + T_o$ with $T_o = 100$.

We consider the four contamination schemes listed in Table 3. Apart from the clean data, which are generated from (5.1) and (5.2), there are three outlier contaminated settings. In the setting with symmetric outliers, the standard deviation of the observation noise σ is replaced by a random variable that has 10% chance of being 10σ and 90% chance of being σ, just as in Section 4.1. In the setting with asymmetric outliers, the mean of the observation noise is replaced by a random variable that has 10% chance of being 10σ and 90% chance of being zero. We expect that this type of outlier will induce an upward bias in the estimation of c, and thus deteriorate forecasting performance of nonrobust estimators. In the setting with bad leverage points, the symmetric outliers have an outlying covariate generated from

$x_t \sim \mathcal{N}(2,100)$ instead of $x_t \sim \mathcal{N}(2,1)$. The outlying covariate makes the observation more influential; we expect this type of outlier to have a larger impact than symmetric outliers. In the out-of-sample period no outliers are present.

We compute the out-of-sample mean squared error, as in (4.2). The average MSE over the 1000 simulations is tabulated in Table 4. The out-of-sample performance is the best using the robust estimators in presence of outliers and is barely worse than the MLE for clean data. The robust estimators have a very similar out-of-sample MSE for outlier contaminated simulations.

Table 4: Out-of-sample MSE for a time varying parameter model, averaged over 1000 simulations.

	MLE	Huber	Trimmed
Clean data	1.04	1.06	1.05
Symmetric outliers	1.22	1.05	1.05
Asymmetric outliers	1.72	1.06	1.06
Bad leverage points	1.43	1.07	1.06

5.2 Real data example

We investigate the effect of personal disposable income (I_t) on personal consumption (C_t) in the United States. The data are quarterly and range from 1959 till the first quarter of 2016. Both time series are plotted in Figure 1a; the units are billions of US dollars. To render them stationary we go in log-differences and get $\Delta \log C_t$ and $\Delta \log I_t$, the series in percentage changes. The scatter plot in Figure 1b suggests a linear relation between these two variables. We consider a time varying parameter model:

$$\Delta \log C_t = c + \theta_t \Delta \log I_t + v_t$$

with θ_t a random walk and v_t the normally distributed error.

We estimate this model with the MLE, the Huber maximum likelihood and the trimmed maximum likelihood estimator. We also fit a linear model, where θ_t is a constant θ, using

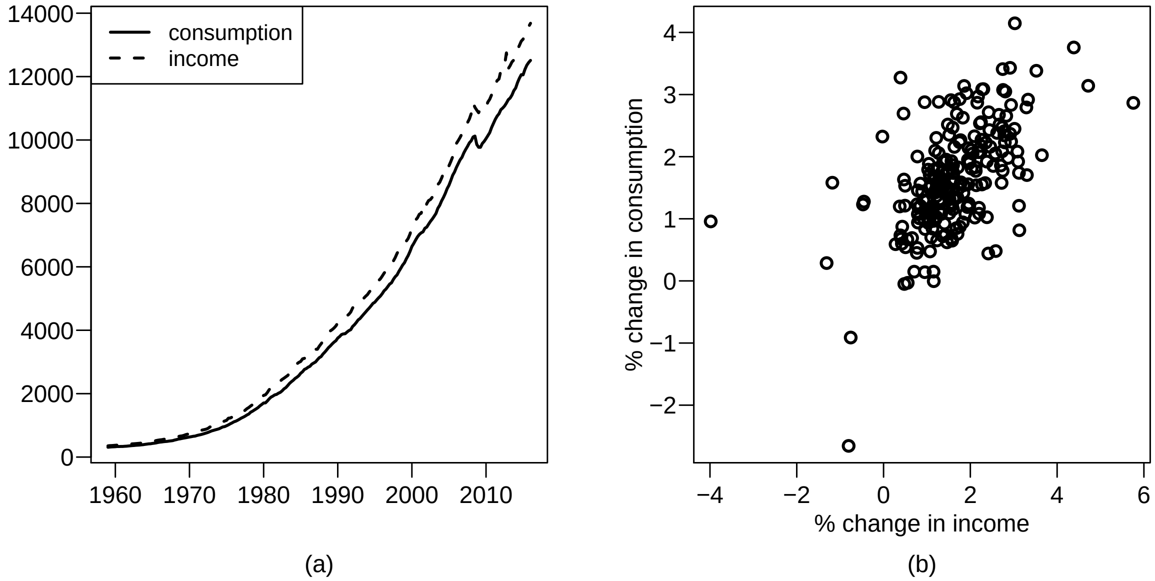

Figure 1: (a) Quarterly time series of personal consumption and income in the US. (b) Scatter plot of personal consumption and income expressed in percentage changes.

the ordinary least squares estimator (OLS) and the robust τ^2 estimator (TAU) of Salibian-Barrera et al. (2008).

The initial values for the optimization routine needed for the estimation of time varying models are the estimates of c and σ with the robust τ^2 estimator. The initial value of λ is equal to 0.1σ, similar as in Stock and Watson (1996).

Each model is estimated using the first $T = 100$ observations. The models are evaluated in an out-of-sample period by computing a mean squared one-step ahead prediction error:

$$\text{MSE} = \frac{1}{128} \sum_{t=101}^{228} (\Delta \log C_t - \Delta \log \hat{C}_{t|t-1}(\hat{\boldsymbol{\phi}}))^2.$$

with $\hat{\boldsymbol{\phi}}$ the estimated parameter vector. The prediction $\Delta \log \hat{C}_{t|t-1}(\hat{\boldsymbol{\phi}})$ makes use of $\Delta \log I_{1:t}$ and $\Delta \log C_{1:t-1}$. In Table 5 this out-of-sample MSE is tabulated. The estimators of the time varying parameter model perform about equally well, and outperform the time invariant models.

Finally, using the Huber estimate of $\boldsymbol{\phi}$, we picture the estimate $\hat{\theta}_{t|t}$ of the time varying coefficient in Figure 2. The 95% confidence interval is $[\hat{\theta}_{t|t} - 1.96\sqrt{\hat{P}_{t|t}}, \hat{\theta}_{t|t} + 1.96\sqrt{\hat{P}_{t|t}}]$,

Table 5: The out-of-sample MSE for predicting consumption growth.

time invariant		time varying		
OLS	TAU	MLE	Huber	Trimmed
$5.54{\cdot}10^{-5}$	$5.31{\cdot}10^{-5}$	$4.15{\cdot}10^{-5}$	$4.11{\cdot}10^{-5}$	$4.33{\cdot}10^{-5}$

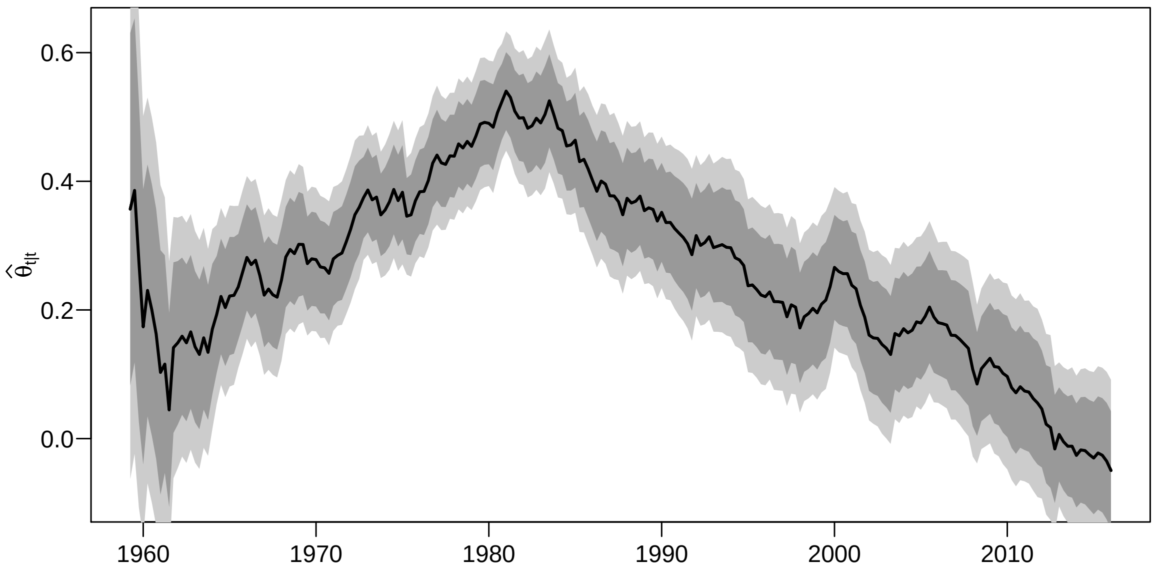

Figure 2: The time varying coefficient $\hat{\theta}_{t|t}$. The dark and light gray zones are the 80% and 95% confidence intervals, respectively.

where $\hat{P}_{t|t}$ is computed with the robust Kalman filter. The coefficient clearly varies over time. As of 1980, the effect of income growth on consumption growth starts to decrease.

6 Conclusion

In earlier papers the main focus is on robust filtering, but not on robust estimation of model parameters of a linear state space model. We robustify the maximum likelihood estimator, combined with a robust filter to approximate the likelihood. The time varying parameter linear regression model is considered as an important special case.

Our proposed method is robust against additive outliers. Such outliers do not persist, and don't contain information about subsequent observations. Other types of outliers, as innovation outliers may occur as well. To have robustness against these outliers we recommend to use a robust filter that is robust against innovation outliers.

In our approach we chose the robust filter of Cipra (1997). There exist other suitable robust filters. If a filter supplies a one-step ahead prediction $\hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})$ and the accompanying one-step ahead prediction error variance $\mathbf{S}_t(\boldsymbol{\phi})$, it can be used instead of the filter of Cipra (1997). We list a number of alternative robust filters. An early suggestion was from Masreliez and Martin (1977). Yang et al. (2001) made an extension and generalization of this method which they call the adaptively robust Kalman filter. This adaptation of the Kalman filter is robust against both additive and innovation outliers. More recently, Gandhi and Mili (2010) proposed the Generalised-Maximum Likelihood Kalman Filter. This filter is also robust against different types of contamination. The disadvantage is that the dimension of the observations needs to be larger than the dimension of the state. Other proposals are in Ruckdeschel et al. (2014) and Marczak et al. (2017).

In this paper we give two proposals to estimate the model parameters robustly: Huber maximum likelihood and maximum trimmed likelihood. These robust estimators produce lower out-of-sample forecasting errors than the Gaussian maximum likelihood estimator if there are outliers. This conclusion is confirmed by the simulations for several settings, including the time varying parameter model.

Acknowledgments. The authors gratefully acknowledge financial support of the KU Leuven grant GOA/12/14.

References

Agamennoni, G., Nieto, J., Nebot, E., 2011. An outlier-robust Kalman filter. IEEE International Conference on Robotics and Automation, 1551–1558.

Brockwell, P., Davis, R., 2002. Introduction to Time Series and Forecasting. Springer, USA.

Cipra, T., 1997. Kalman filter with outliers and missing observations. Test 6 (2), 379–395.

Croux, C., Haesbroeck, G., 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. Journal of Multivariate Analysis 71 (2), 161 – 190.

Durbin, J., Koopman, S., 2012. Time Series Analysis by State Space Methods. Oxford University Press, UK.

Gandhi, M., Mili, L., 2010. Robust Kalman filter based on a generalized maximum-likelihood-type estimator. IEEE Transactions on Signal Processing 58 (5), 2509–2520.

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

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

Marczak, M., Proietti, T., Grassi, S., 2017. A data-cleaning augmented Kalman filter for robust estimation of state space models. Econometrics and Statistics, 2452–3062.

Masreliez, C., Martin, R., 1977. Robust bayesian estimation for the linear model and robustifying the Kalman filter. IEEE transactions on Automatic Control 22 (3), 361–371.

Muler, N., Peña, D., Yohai, V. J., 2009. Robust estimation for ARMA models. The Annals of Statistics 37 (2), 816–840.

Petris, G., Petrone, S., Campagnoli, P., 2009. Dynamic Linear Models with R. Springer.

Rosenberg, B., 1972. The estimation of stochastic regression parameters re-examined. Journal of the American Statistical Association 67 (339), 650–654.

Ruckdeschel, P., Spangl, B., Pupashenko, D., 2014. Robust Kalman tracking and smoothing with propagating and non-propagating outliers. Statistical Papers 55, 93–123.

Salibian-Barrera, M., Willems, G., Zamar, R., 2008. The fast-τ estimator for regression. Journal of Computational and Graphical Statistics 17 (3), 659–682.

Stock, J., Watson, M., 1996. Evidence on structural instability in macroeconomic time series relations. Journal of Business and Economic Statistics 14 (1), 11–30.

Yang, Y., He, H., Xu, G., 2001. Adaptively robust filtering for kinematic geodetic positioning. Journal of Geodesy 75, 109–116.

A Derivation of filtered state estimates through least squares

We prove equation (3.1)

$$\hat{\boldsymbol{\theta}}_{t|t} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \left(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}\right)' \mathbf{P}_{t|t-1}^{-1} \left(\hat{\boldsymbol{\theta}}_{t|t-1} - \boldsymbol{\theta}\right) + \left(\mathbf{y}_t - \mathbf{H}_t\boldsymbol{\theta}\right)' \boldsymbol{\Sigma}_t^{-1} \left(\mathbf{y}_t - \mathbf{H}_t\boldsymbol{\theta}\right).$$

The derivative with respect to $\boldsymbol{\theta}$ is

$$2\mathbf{P}_{t|t-1}^{-1} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{t|t-1}\right) - 2\mathbf{H}_t'\boldsymbol{\Sigma}_t^{-1} \left(\mathbf{y}_t - \mathbf{H}_t\boldsymbol{\theta}\right).$$

Because $\mathbf{P}_{t|t-1}$ and $\boldsymbol{\Sigma}_t$ are positive definite, setting the derivative equal to zero and solving for $\boldsymbol{\theta}$ gives the solution:

$$\begin{aligned}
\boldsymbol{\theta} &= \left(\mathbf{I} + \mathbf{P}_{t|t-1}\mathbf{H}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{H}_t\right)^{-1} \left(\mathbf{P}_{t|t-1}\mathbf{H}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{Y}_t + \hat{\boldsymbol{\theta}}_{t|t-1}\right) \\
&= \left(\mathbf{P}_{t|t-1}^{-1} + \mathbf{H}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{H}_t\right)^{-1} \left(\mathbf{H}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{Y}_t + \mathbf{P}_{t|t-1}^{-1}\hat{\boldsymbol{\theta}}_{t|t-1}\right).
\end{aligned}$$

Using the Woodbury matrix identity we find:

$$\begin{aligned}
\boldsymbol{\theta} &= \left(\mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}_t'\left(\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \boldsymbol{\Sigma}_t\right)^{-1}\mathbf{H}_t\mathbf{P}_{t|t-1}\right)\left(\mathbf{H}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{Y}_t + \mathbf{P}_{t|t-1}^{-1}\hat{\boldsymbol{\theta}}_{t|t-1}\right) \\
&= \hat{\boldsymbol{\theta}}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}_t'\left(\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \boldsymbol{\Sigma}_t\right)^{-1}\mathbf{H}_t\hat{\boldsymbol{\theta}}_{t|t-1} \\
&\quad + \mathbf{P}_{t|t-1}\mathbf{H}_t'\left(\mathbf{I} - \left(\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \boldsymbol{\Sigma}_t\right)^{-1}\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t'\right)\boldsymbol{\Sigma}_t^{-1}\mathbf{Y}_t \\
&= \hat{\boldsymbol{\theta}}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}_t'\left(\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \boldsymbol{\Sigma}_t\right)^{-1}\mathbf{H}_t\hat{\boldsymbol{\theta}}_{t|t-1} \\
&\quad + \mathbf{P}_{t|t-1}\mathbf{H}_t'\left(\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \boldsymbol{\Sigma}_t\right)^{-1}\mathbf{Y}_t \\
&= \hat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{H}_t'\left(\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t' + \boldsymbol{\Sigma}_t\right)^{-1}\left(\mathbf{Y}_t - \mathbf{H}_t\hat{\boldsymbol{\theta}}_{t|t-1}\right) \\
&= \hat{\boldsymbol{\theta}}_{t|t-1} + \mathbf{K}_t\left(\mathbf{Y}_t - \mathbf{H}_t\hat{\boldsymbol{\theta}}_{t|t-1}\right),
\end{aligned}$$

which is equal to $\hat{\boldsymbol{\theta}}_{t|t}$ as given in the Kalman recursions (2.2).

B Computation of constants c_H and c_T

The constants c_H and c_T make that the expected value of respectively the Huber likelihood in (3.6) and the trimmed likelihood in (3.8) are the same as the expected value of the likelihood in (2.3) if $\boldsymbol{\phi}$ is the true model parameter, and if the Kalman filter is used. The constants can be computed analytically. The standardized residual $\mathbf{x} = \mathbf{S}_t(\boldsymbol{\phi})^{-1/2}\left(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}(\boldsymbol{\phi})\right)$ has a multivariate standard normal distribution. Therefore its norm has a χ_d distribution. We need to set c_H such that

$$c_H = \frac{\frac{1}{2}\mathbb{E}(X^2)}{\mathbb{E}(\rho_H(X))}$$

where X has a χ_d distribution with d degrees of freedom. From (3.7) we have

$$\mathbb{E}(\rho_H(X)) = \frac{1}{2}\mathbb{E}(X^2\ I(X < k)) + \mathbb{E}((kX - \frac{k^2}{2})\ I(X > k))$$

$Z = X^2$ has a χ_d^2 distribution. We find that

$$\begin{aligned}
\mathbb{E}(X^2\ I(X^2 < k^2)) &= \mathbb{E}(Z\ I(Z < k^2)) \\
&= \int_0^{k^2} z \frac{1}{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)} z^{\frac{d}{2}-1} \exp\left(-\frac{z}{2}\right) dz \\
&= \int_0^{k^2} \frac{1}{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)} z^{\frac{d+2}{2}-1} \exp\left(-\frac{z}{2}\right) dz \qquad \text{(B.1)} \\
&= d\int_0^{k^2} \frac{1}{2^{\frac{d+2}{2}}\Gamma\left(\frac{d+2}{2}\right)} z^{\frac{d+2}{2}-1} \exp\left(-\frac{z}{2}\right) dz \\
&= d\ F_{\chi^2_{d+2}}(k^2)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(X\ I(X > k)) &= \int_k^{+\infty} x \frac{1}{2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)} x^{d-1} \exp\left(-\frac{x^2}{2}\right) dx \\
&= \int_k^{+\infty} \frac{1}{2^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)} x^{(d+1)-1} \exp\left(-\frac{x^2}{2}\right) dx \\
&= \sqrt{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \int_k^{+\infty} \frac{1}{2^{\frac{d+1}{2}-1}\Gamma\left(\frac{d+1}{2}\right)} x^{(d+1)-1} \exp\left(-\frac{x^2}{2}\right) dx \\
&= \sqrt{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}(1 - F_{\chi_{d+1}}(k)) = \sqrt{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}(1 - F_{\chi^2_{d+1}}(k^2)).
\end{aligned}$$

Since $\mathbb{E}(X^2) = d$, we find

$$c_H = \frac{d}{dF_{\chi^2_{d+2}}(k^2) + 2k\sqrt{2}\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}(1 - F_{\chi^2_{d+1}}(k^2)) - k^2(1 - F_{\chi^2_d}(k^2))}$$

with $F_{\chi^2_d}$ is the cumulative distribution function of a χ^2 distribution with d degrees of freedom.

For the maximum trimmed likelihood, the constant c_T needs to be set at

$$c_T = \frac{\mathbb{E}(X^2)}{\mathbb{E}(X^2\ I(X^2 < F^{-1}_{\chi^2_d}(1 - \alpha)))},$$

with $F_{\chi_d^2}^{-1}$ the quantile function of a χ^2 distribution with d degrees of freedom. By setting $k^2 = F_{\chi_d^2}^{-1}(1-\alpha)$ in (B.1), we get

$$c_T = \frac{1}{F_{\chi_{d+2}^2}(F_{\chi_d^2}^{-1}(1-\alpha))}.$$

This constant is equal to the consistency factor of the minimum covariance determinant estimator, computed in Croux and Haesbroeck (1999).

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be