# Robust filtering

Calvet, Laurent; Czellar, Veronika; Ronchetti, Elvezio

# Robust Filtering

**Laurent E. Calvet, Veronika Czellar and Elvezio Ronchetti**[*]

First version: July 2011
This version: June 11, 2013

[*] Calvet: Department of Finance, HEC Paris, 1 rue de la Libération, 78351 Jouy en Josas, France; calvet@hec.fr. Czellar: Department of Economics and Decision Sciences, HEC Paris, 1 rue de la Libération, 78351 Jouy en Josas, France, czellarv@hec.fr. Ronchetti: Research Center for Statistics and Department of Economics, University of Geneva, Blv. Pont d'Arve 40 CH-1211 Geneva, Switzerland, elvezio.ronchetti@unige.ch. The authors thank the Editor, the Associate Editor, and a Referee for useful comments which improved the original version of the manuscript.

### Abstract

Filtering methods are powerful tools to estimate the hidden state of a state-space model from observations available in real time. However, they are known to be highly sensitive to the presence of small misspecifications of the underlying model and to outliers in the observation process. In this paper, we show that the methodology of robust statistics can be adapted to sequential filtering. We introduce an impact function that quantifies the sensitivity of the state distribution with respect to new data. Since the impact function of standard filters are unbounded even in the simplest cases, we propose filters with bounded impact functions which provide accurate state and parameter inference in the presence of model misspecifications. In particular, the robust particle filter naturally solves the degeneracy problems that plague the bootstrap particle filter (Gordon, Salmond and Smith, 1993) and its many extensions. We illustrate the good properties of robust filters in several examples, including linear state-space models and nonlinear models of stochastic volatility.

**Keywords:** Kalman filter, Kullback-Leibler divergence, particle filter, robust statistics, state-space model, stochastic volatility, weight degeneracy.

# 1 Introduction

Filtering techniques are used in many different fields to sequentially estimate the hidden states of a state-space model from data observed in real time. More specifically, suppose that at date $t \in \mathbb{N}$, a set of observations $Y_t = \{y_1, \ldots, y_t\}$, $y_t \in \mathbb{R}^p$, has been collected, and assume that they have been generated from a general state-space model specified by a Markov process $x_t$ with *kernel* $\rho(x_t|x_{t-1})$ and *observation density* $f(y_t|x_t, Y_{t-1})$. Examples include linear state-space models (see e.g. Harvey, 1989) and nonlinear systems of the type commonly used in engineering and finance (see Section 4).

The statistician is interested in estimating the *filtering distribution* $\lambda(x_t|Y_t)$ of the states, as well as the unknown parameters of the kernel and the observation density. The estimation is usually based on Bayes' rule:

$$\lambda(x_t|y_t, Y_{t-1}) \propto f(y_t|x_t, Y_{t-1})\lambda(x_t|Y_{t-1}). \tag{1.1}$$

The filtering distribution has an analytical expression in selected cases, such as linear Gaussian (Kalman, 1960) or finite state-space models (Hamilton, 1989; Lindgren, 1978). In more complex situations, implementation of (1.1) can proceed by particle filter approximations and related methods (Gordon, Salmond and Smith, 1993), for which a large body of literature exists. Good textbook treatments include Del Moral (2004) and Cappé, Moulines and Rydén (2005), and useful overviews are Doucet, De Freitas and Gordon (2001) (basic introduction), Doucet and Johansen (2011) (unified framework and recent results) and Johannes and Polson (2009) (finance focus).

While the Bayesian filter and its refinements are very powerful techniques which flexibly adapt to new incoming data, they are also highly sensitive to outliers in the observation process. To illustrate this point, suppose that at a given date $t$, instead of $y_t$ we observe a noisy version $y_t^{cont}$ of the form

$$y_t^{cont} = y_t + v_t \,, \tag{1.2}$$

where the contaminating process $v_t \in \mathcal{V}_k(y_t; Y_{t-1})$ is unknown (see Assumption 1). This includes contaminations such as additive outliers (AO), replacement outliers (RO) and innovative outliers (IO) typically considered in the robustness literature; see Maronna, Martin and Yohai (2006) p. 252 ff. Since the filtered distribution at date $t$ is calculated using the structural non-contaminated observation density $f(y_t^{cont}|x_t, Y_{t-1})$ evaluated at $y_t^{cont}$ instead of the unavailable contaminated observation density $f_{cont}(y_t^{cont}|x_t, Y_{t-1})$ in (1.2), we can ask about the difference between the former and the latter. This quantifies the impact of the contamination on the estimation of the filtering distribution.

To measure the accuracy of the estimation, we use the backward Kullback-Leibler divergence between the structural and exact filtering distributions:

$$KL_t = KL\big[\lambda(x_t|y_t^{cont}, Y_{t-1}), \lambda_{cont}(x_t|y_t^{cont}, Y_{t-1})\big], \tag{1.3}$$

where $\lambda_{cont}(x_t|y_t^{cont}, Y_{t-1}) \propto f_{cont}(y_t^{cont}|x_t, Y_{t-1})\lambda_{cont}(x_t|Y_{t-1})$ and $\lambda_{cont}(x_t|Y_{t-1})$ coincides with $\lambda(x_t|Y_{t-1})$ if no outliers occurred prior to date $t$. As an illustrative example, the top panel of Figure 1 reports observations generated from a Markov-switching multifractal (MSM) volatility model without contamination (left panel) and under 5% contamination (right panel, where 5% of randomly chosen observations are magnified by a factor of 4). The exact specification of the MSM model along with the parameter choices are provided in Section 4.1.1. The middle panel of

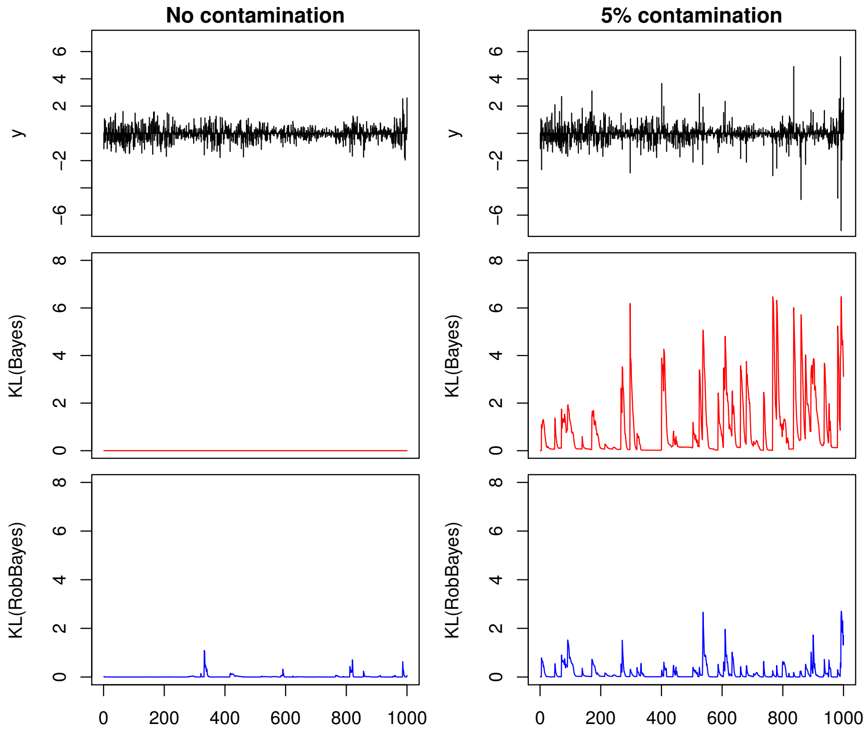

Figure 1: Accuracy of standard and robust filters for an MSM volatility model without contamination (left panel) and with 5% contamination (right panel).

Figure 1 reports the Kullback-Leibler divergence $KL_t$, $t = 1, \ldots, T$, in (1.3) for the standard Bayesian filter. While in the uncontaminated case (left panel) the standard Bayesian filter shows good accuracy, under a small amount of contamination (right panel) it becomes very inaccurate. This example illustrates that the Bayesian filter loses its power to extract the underlying state in the presence of outliers in the observation process.

The bottom panel of Figure 1 reports the accuracy of a robust filter that will be introduced in Section 3 of this paper. The robust filter performs like its standard counterpart in the uncontaminated case. It can also withstand small misspecifications in the observation model and exhibit very good accuracy under contamination.

Robust statistics deals with deviations from parametric models and develops inference procedures which are not unduly influenced by such deviations. A large body of literature exists and robust estimators and tests are available for many different models; see the classical books by Huber (1981, 2nd edition by Huber and Ronchetti 2009), Hampel et al. (1986), and Maronna, Martin and Yohai (2006). In particular, robust alternatives to the Kalman filter have been proposed; see in particular Masreliez and Martin (1977), the good overview in Schick and Mitter (1994) with the references thereof, and more recent work by Ruckdeschel (2010a), Ruckdeschel (2010b), Ruckdeschel, Spangl and Pupashenko (2012). A robust Kalman filter can be viewed as a special case of our framework, since it provides the expected value of the filtering distribution. In this paper we go beyond the basic framework of the Kalman filter by studying the robustness properties of the entire estimated filtering distribution and by deriving a robust particle filter which is a robust estimator of the latter.

More specifically, our contributions to the literature are as follows. In Section 2, we define a filter's impact function, which measures how the innovation $y_t$ impacts the state distribution of $x_t$ by means of the backward Kullback-Leibler divergence between $\lambda(x_t|Y_{t-1})$ and $\lambda(x_t|y_t, Y_{t-1})$. By requiring a bounded impact function, we can state a sufficient condition for the robustness of a filter. It is based on the observation density and can be used as a robustness diagnostic tool for any specific model.

In Section 3, we obtain a general robust filter by appropriately truncating the observation density, which is in line with the basic principles of robust statistics. The construction is achieved by "huberizing" the derivative of the log-observation density $(\partial \log f)/(\partial y)$ and then by integrating it. When the original observation density is a (univariate or multivariate) Gaussian, the robustified observation density has a closed-form expression and efficiency provides a natural selection method for the tuning constant, which gives guidelines to researchers for the operational development of the filter. Furthermore, for some special systems, we develop an alternative robustification method based on the Student $t$ distribution and derive a selection criterion for the number of degrees of freedom of the robust Student $t$ filter. In applications, robust filtering can be applied in closed form when the state space is finite, or by way of a particle filter in richer environments. One key advantage of the robust approach is that it naturally overcomes the degeneracy problem that plagues the bootstrap filter and its many extensions (Gordon, Salmond and Smith, 1993; Pitt and Shephard, 1999).

In Section 4, we show the generality of our results by presenting three main applications. The first example is a Markov-Switching Multifractal model (Calvet and Fisher, 2001), a complex model for which classical estimation is non-trivial and no robust procedures are available. In addition to the improvements in filtering reported in Figure 1, our procedure is shown to permit robust model selection, applied to the number of volatility factors. The second example is a linear Gaussian state-space model. We summarize available robust Kalman filter procedures and compare their performance with that of the standard Kalman filter and our robust particle filter. Although the new robust particle filter is not tuned for a linear system, we show that under contamination its performance is comparable to the benchmark robust Kalman filters. Of course both clearly outperform standard techniques in the presence of contamination. The third example is a stochastic volatility model. We show that the robust filter solves the degeneracy problem of classical particle filters, as Figure 9 illustrates, and accurately tracks the underlying state under contamination. In contrast to standard methods, the robust filter generates a simulated likelihood function that varies smoothly with the parameters of the model. Furthermore, the robust filter greatly reduces the bias exhibited by the standard maximum likelihood estimator of the model parameters in the presence of even small amounts of contamination.

Section 5 concludes with possible directions for future research. Assumptions and proofs are provided in the Appendix.

# 2 A Robustness Measure for the Filtering Distribution

## 2.1 Impact Function

We measure the impact of the innovation $y_t$ on the conditional distribution of $x_t$ by the backward Kullback-Leibler divergence between $\lambda(x_t|Y_{t-1})$ and $\lambda(x_t|y_t, Y_{t-1})$:

$$KL\big[\lambda(x_t|Y_{t-1}), \lambda(x_t|y_t, Y_{t-1})\big] = \mathbb{E}_{\lambda(x_t|Y_{t-1})}\left[\log \frac{\lambda(x_t|Y_{t-1})}{\lambda(x_t|y_t, Y_{t-1})}\right] . \qquad (2.1)$$

Assume that at date $t$, the contaminated data point $y_t^{cont} = y_t + v_t$ is observed instead of $y_t$. An expansion of the backward Kullback-Leibler divergence around $v_t = 0$ gives

$$\begin{aligned} KL\big[\lambda(x_t|Y_{t-1}), \lambda(x_t|y_t + v_t, Y_{t-1})\big] &= KL\big[\lambda(x_t|Y_{t-1}), \lambda(x_t|y_t, Y_{t-1})\big] + \\ &+ v_t' \frac{\partial}{\partial y_t} KL\big[\lambda(x_t|Y_{t-1}), \lambda(x_t|y_t, Y_{t-1})\big] + \mathcal{O}(\|v_t\|^2) . \end{aligned}$$

This leads to the following definition.

**Definition 1 (Impact function)** *Under contamination (1.2), the impact function* *at date $t$ of $\lambda(x_t|y_t, Y_{t-1})$ is defined by*

$$I(y_t; \lambda, Y_{t-1}, v_t) = v_t' \frac{\partial}{\partial y_t} KL\big[\lambda(x_t|Y_{t-1}), \lambda(x_t|y_t, Y_{t-1})\big] \tag{2.2}$$

*for every $v_t \in \mathbb{R}^p$.*

The impact function quantifies the sensitivity of the filtered distribution with respect to the contaminated observation. For this reason, it is the filtering analogue of the sensitivity curve that is commonly considered in robust statistics.

Bayes' rule allows us to relate the impact function to the sensitivity of the observation density.

**Proposition 1 (Analytical expression of the impact function)** *Under the contamination (1.2), the impact function is given by*

$$I(y_t; \lambda, Y_{t-1}, v_t) = v_t' \left\{ \mathbb{E}_{\lambda(x_t|Y_t)} \left[ \frac{\partial \log f(y_t|x_t, Y_{t-1})}{\partial y_t} \right] - \mathbb{E}_{\lambda(x_t|Y_{t-1})} \left[ \frac{\partial \log f(y_t|x_t, Y_{t-1})}{\partial y_t} \right] \right\}$$

*for every $v_t \in \mathbb{R}^p$.*

*Illustrative Example.* Consider a state-space model where the observation density $f(y_t|x_t, Y_{t-1})$ is univariate Gaussian $\mathcal{N}[0, \sigma(x_t)^2]$. By Proposition 1, the impact function at date $t$ is

$$I(y_t; \lambda, Y_{t-1}, v_t) = \left\{ \mathbb{E}[\sigma(x_t)^{-2}|Y_{t-1}] - \mathbb{E}[\sigma(x_t)^{-2}|Y_t] \right\} y_t v_t \,.$$

Suppose that the state space is finite and that the transition probabilities between the states are all strictly positive (as in the case of MSM volatility models, see

Section 4.1.1). Let $u = \max_{x_t} \sigma(x_t)$. Then, $\lim_{|y_t|\to\infty} \mathbb{E}[\sigma(x_t)^{-2}|Y_t] = u^{-2}$ and as $|y_t| \to \infty$ the impact function is equivalent to

$$\left\{\mathbb{E}[\sigma(x_t)^{-2}|Y_{t-1}] - u^{-2}\right\} y_t v_t \,.$$

The impact function $I(y_t; \lambda, Y_{t-1}, v_t)$ is asymptotically proportional to $y_t$ and can therefore be arbitrarily large. The next subsection provides a general sufficient condition guaranteeing the boundedness of the impact function under an appropriate class of disturbances.

## 2.2 A Robustness Condition

If the impact function (2.2) has a bounded linear coefficient $\partial KL/\partial y_t$, the filtering distribution cannot be driven arbitrarily away by a very small contamination. This motivates the following definition of robustness.

**Definition 2 (Robustness of $\lambda$)** *Let $\mathcal{V}_k(y_t; Y_{t-1})$ denote the class of admissible disturbances defined in Assumption 1. The filtering distribution $\lambda(x_t|Y_t)$ is robust with respect to $\mathcal{V}_k(y_t; Y_{t-1})$, if there exists a positive constant $\tilde{c}$ such that*

$$|I(y_t; \lambda, Y_{t-1}, v_t)| \leq \tilde{c}\,.$$

*for every $y_t \in \mathbb{R}^p$ and $v_t \in \mathcal{V}_k(y_t, Y_{t-1})$.*

Robustness is guaranteed by the following key criterion.

**Proposition 2 (Sufficient condition for robustness)** *If Assumptions 1 and 2* *hold and there exists $c \in \mathbb{R}_+$ such that for all $x_t, y_t$*

$$\left\| \frac{\partial \log f(y_t|x_t, Y_{t-1})}{\partial y_t} \right\| \|y_t - \mathbb{E}(y_t|Y_{t-1})\| \leq c\,, \tag{2.3}$$

*then the filter is robust.*

Condition (2.3) can be used as a diagnostic tool to check the robustness of a given filter. For instance, the Bayesian filter in the illustrative example of Section 2.1 fails to meet (2.3) for any choice of $c \in \mathbb{R}_+$. In the next section we provide an explicit construction of a robust filter for a general state-space model.

# 3 Robust Filters

## 3.1 A General $C^1$ Solution

We develop a general robust solution to (2.3), which has the key feature of being continuously differentiable on the observation space. The $C^1$ property will imply that robustified procedures have excellent numerical stability, as will be illustrated in Section 4.

The construction consists of "huberizing" the derivative of the log-observation density $(\partial \log f)/(\partial y_t)$ and then integrating it to obtain the robust density. Let

$$\mu(x_t) = \mathbb{E}(y_t|x_t, Y_{t-1}),$$
$$\mu_t = \mathbb{E}(y_t|Y_{t-1}).$$

We show in the Appendix the following result.

**Proposition 3 (Nonnormalized robustified observation density)** *We consider that Assumptions 2 and 3 hold and define*

$$g(y) = h_{\frac{c}{\|y-\mu_t\|}} \left[ \frac{\partial \log f}{\partial y}(y|x_t, Y_{t-1}) \right], \tag{3.1}$$

*where* $h_\tau(z) = z \min(1; \tau/\|z\|)$ *is the multivariate Huber function and* $c \in \mathbb{R}_+$ *is a tuning constant. Then the function*

$$\tilde{f}(y_t|x_t, Y_{t-1}) = f[\mu(x_t)|x_t, Y_{t-1}] \exp\left\{ \int_0^1 [y_t - \mu(x_t)]' \, g[\mu(x_t) + s(y_t - \mu(x_t))] \, ds \right\} \tag{3.2}$$

*belongs to* $C^1(\mathbb{R}^p)$ *(i.e. is continuously differentiable everywhere) and satisfies the robustness condition (2.3) for every* $y_t \in \mathbb{R}^p$*.*

The solution $\tilde{f}$ coincides with the observation density if the tuning constant is infinite. More generally, a high tuning constant corresponds to mild robustification, while a low constant implies strong robustification.

The function $\tilde{f}(y_t|x_t, Y_{t-1})$ in (3.2) generally does not integrate to unity and must be normalized to obtain a proper density. Consider the *normalized robustified observation density*

$$\hat{f}(y_t|x_t, Y_{t-1}) = B_t(x_t)\tilde{f}(y_t|x_t, Y_{t-1}) \tag{3.3}$$

with $B_t(x_t)$ such that $\int_{\mathbb{R}^p} B_t(x_t)\tilde{f}(y_t|x_t, Y_{t-1})dy_t = 1$. Assume that conditional on $Y_{t-1}$, the state $x_t$ is drawn from $\lambda(x_t|Y_{t-1})$. We then define the *robustified one-step likelihood* associated with the normalized robustified density by

$$\hat{f}(y_t|Y_{t-1}) = \mathbb{E}_{\lambda(x_t|Y_{t-1})}[\hat{f}(y_t|x_t, Y_{t-1})] \, . \tag{3.4}$$

In Section 3, we will introduce a robustified version of the full likelihood of $Y_T$.

Huberizing the derivative of the observation density ensures robustness to outliers, but this comes at the cost of a loss of accuracy in the likelihood approximation when $f(y_t|Y_{t-1})$ is the true data-generating process of $Y_t$. We measure the *efficiency cost* by the Kullback-Leibler divergence:

$$KL_t^{\text{eff}} = KL[f(y_t|Y_{t-1}), \hat{f}(y_t|Y_{t-1})] = \mathbb{E}_{f(y_t|Y_{t-1})}\left\{\log\left[\frac{f(y_t|Y_{t-1})}{\hat{f}(y_t|Y_{t-1})}\right]\right\}. \tag{3.5}$$

The analyst can choose the tuning constant $c$ such that the efficiency cost $KL_t^{\text{eff}}$ is always less than a given maximal deviation $\alpha$ (typically 1% or 5%). Specific formulas for the upper bound of $KL_t^{\text{eff}}$ are provided in Section 3.4, and a method to estimate $KL_t^{\text{eff}}$ is described in Section 4.3.

## 3.2 Robustifying a Univariate Gaussian Model

In this section, we consider that the observation density $f(y_t|x_t, Y_{t-1})$ is a univariate Gaussian with mean $\mu(x_t)$ and variance $\sigma^2(x_t)$, where $\mu(x_t)$ and $\sigma^2(x_t)$ may depend on $Y_{t-1}$. Condition (2.3) holds as an equality if and only if

$$|y_t - \mu(x_t)|\ |y_t - \mu_t| = c\,\sigma^2(x_t). \tag{3.6}$$

This equation has two distinct roots $y_t$ if $c > [\mu(x_t) - \mu_t]^2/[2\sigma(x_t)]^2$, and four distinct roots otherwise. We separately examine these two cases.

First, assume that $c > [\mu(x_t) - \mu_t]^2/[2\sigma(x_t)]^2$. This condition holds for instance if $\mu(x_t) = \mu_t$, or if $\mu(x_t) \neq \mu_t$ and robustification is mild. Equation (3.6) then has two distinct roots:

$$y_\pm^* = \frac{\mu(x_t) + \mu_t \pm \sqrt{(\mu(x_t) - \mu_t)^2 + 4c\sigma(x_t)^2}}{2}. \tag{3.7}$$

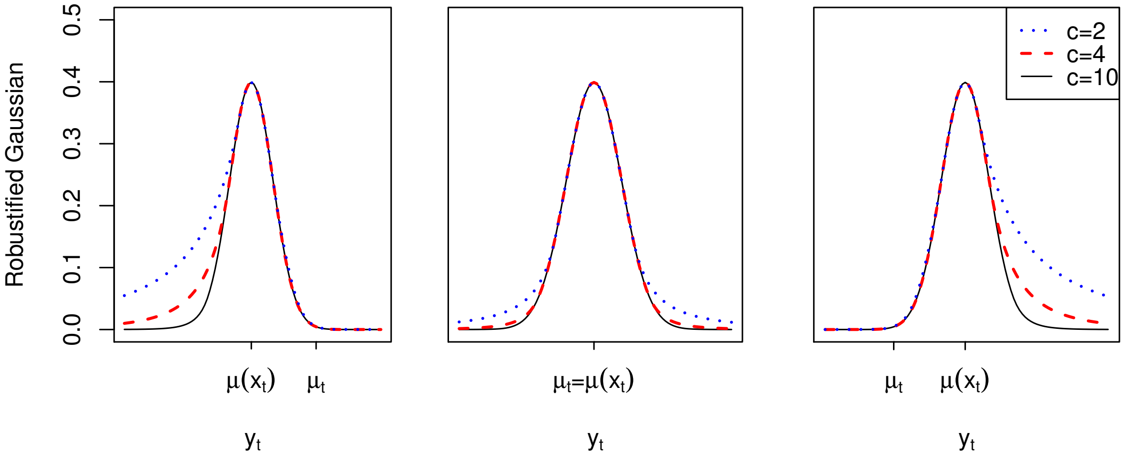

Figure 2: Robustified Gaussian. The figure illustrates the nonnormalized robustified Gaussian $\tilde{f}(y_t|x_t, Y_{t-1})$ defined by (3.9).

The nonnormalized robustified density (3.2) is

$$\tilde{f}(y_t|x_t, Y_{t-1}) = \begin{cases} D_{1,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t < y_-^*, \\ f_{\mathcal{N}}[y_t; \mu(x_t), \sigma(x_t)^2] & \text{if } y_t \in [y_-^*, y_+^*), \\ D_{2,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t \geq y_+^*, \end{cases} \tag{3.8}$$

where $f_{\mathcal{N}}[\cdot; \mu(x_t), \sigma(x_t)^2]$ is the density of the normal distribution with mean $\mu(x_t)$ and variance $\sigma(x_t)^2$, and $D_{1,t}(x_t)$ and $D_{2,t}(x_t)$ are chosen to guarantee continuity at $y_-^*$ and $y_+^*$ (see Appendix for explicit formulas.)

Conversely, if $c \leq [\mu(x_t) - \mu_t]^2/[2\sigma(x_t)]^2$, condition (3.6) holds for the values $y_-^*$ and $y_+^*$ defined by the maintained equation (3.7), as well as for the two additional roots:

$$z_\pm^* = \frac{\mu(x_t) + \mu_t \pm \sqrt{(\mu(x_t) - \mu_t)^2 - 4c\sigma(x_t)^2}}{2}.$$

We observe that $y_-^* < z_-^* \leq z_+^* < y_+^*$. If $\mu(x_t) \leq \mu_t$, the nonnormalized robustified density satisfies:

$$\tilde{f}(y_t|x_t, Y_{t-1}) = \begin{cases} C_{1,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t < y_-^* \\ f_{\mathcal{N}}[y_t; \mu(x_t), \sigma(x_t)^2] & \text{if } y_t \in [y_-^*, z_-^*) \\ C_{2,t}(x_t)|y_t - \mu_t|^{c} & \text{if } y_t \in [z_-^*, z_+^*) \\ C_{3,t}(x_t) f_{\mathcal{N}}[y_t; \mu(x_t), \sigma(x_t)^2] & \text{if } y_t \in [z_+^*, y_+^*) \\ C_{4,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t \geq y_+^* \end{cases} . \tag{3.9}$$

The constants $C_{1,t}(x_t), \ldots, C_{4,t}(x_t)$ are chosen to guarantee the continuity of the robustified density and are provided in closed form in the Appendix. A similar definition holds if $\mu(x_t) > \mu_t$. The robustified Gaussian is plotted in Figure 2 for different values of $c$. We will show in Section 3.4 that the normalizing constant $B_t(x_t)$ in (3.3) has an explicit form and can be used to define a selection rule for the tuning constant $c$.

## 3.3 Robustifying a Multivariate Gaussian Model

We now assume that $y_t \in \mathbb{R}^p$ and that the observation density $f(y_t|x_t, Y_{t-1})$ is a multivariate Gaussian with mean $\mu(x_t)$ and variance-covariance matrix $\Sigma(x_t)$, denoted $f_{\mathcal{N}}[y_t; \mu(x_t), \Sigma(x_t)]$. Despite the conveniently short notation, the mean $\mu(x_t)$ and variance $\Sigma(x_t)$ can also depend on $Y_{t-1}$. The robustified density (3.2) is defined as a line integral over the segment $[\mu(x_t), y_t]$, which we can subdivide into truncation and no-truncation subsegments. For example, if $p = 2$ and the Gaussian is spherical: $\Sigma(x_t) = \sigma^2(x_t)\, I_{2\times 2}$, the no-truncation region: $\{y \in \mathbb{R}^p : \|y - \mu(x_t)\|\, \|y - \mu_t\| \leq c\sigma^2(x_t)\}$ is bounded by a Cassini oval with foci $\mu_t$ and $\mu(x_t)$ (see, e.g., Lockwood,

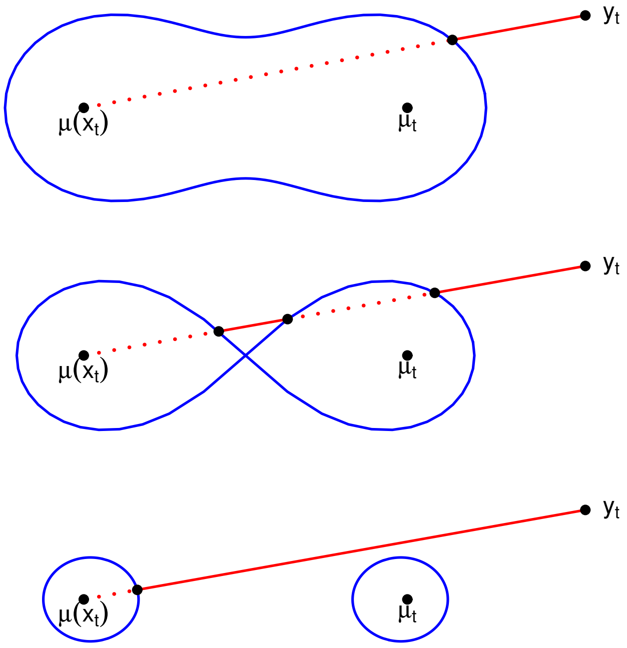

Figure 3: Cassini ovals for three different values of $c$. The segment between $\mu(x_t)$ and $y_t$ corresponds to the integration domain in Proposition 4. Dotted lines indicate the no-truncation region and continuous lines the truncation region.

1961), as Figure 3 illustrates. The segment $[\mu(x_t), y_t]$ can intersect the Cassini oval multiple times, and is correspondingly partitioned into truncation (dotted lines) and no-truncation (solid lines) subsegments.

Suppose that the segment $[\mu(x_t), y_t]$ in $\mathbb{R}^p$ intersects the boundary of the non-truncation region (3.1) $J \geq 0$ times in the values $\{y_t^{(j)}\}_{j=1,\ldots,J}$. We let $y_t^{(0)} = \mu(x_t)$ and $y_t^{(J+1)} = y_t$. Partition the unit interval into $s_0 = 0 < s_1 < \ldots < s_J < s_{J+1} = 1$ such that

$$y_t^{(j)} = \mu(x_t) + s_j[y_t - \mu(x_t)] \,, \quad j = 0, 1, \ldots, J+1 \,. \tag{3.10}$$

Then, by subdividing the integral in (3.2) in $J+1$ parts delimited by $\{s_j\}_{j=0,\ldots,J+1}$ we obtain the robustified multivariate Gaussian summarized in the following proposition.

**Proposition 4 (Robustified multivariate Gaussian)** *Given a multivariate normal observation density $f_{\mathrm{N}}[y_t; \mu(x_t), \Sigma(x_t)]$ on $\mathbb{R}^p$, the robustified density is*

$$\tilde{f}(y_t|x_t, Y_{t-1}) = f_{\mathrm{N}}[y_t^{(1)}; \mu(x_t), \Sigma(x_t)] \prod_{\substack{j=1 \\ j \text{ even}}}^{J} \frac{f_{\mathrm{N}}[y_t^{(j+1)}; \mu(x_t), \Sigma(x_t)]}{f_{\mathrm{N}}[y_t^{(j)}; \mu(x_t), \Sigma(x_t)]} \prod_{\substack{j=1 \\ j \text{ odd}}}^{J} \frac{q[y_t^{(j+1)}; \mu_t, x_t]}{q[y_t^{(j)}; \mu_t, x_t]}$$

*where $\{y_t^{(j)}\}_{j=1,\ldots,J+1}$ are defined in (3.10) and*

$$q(y; \mu_t, x_t) = \left\{ \|y - \mu_t\| + \frac{[y - \mu(x_t)]'(y - \mu_t)}{\|y - \mu(x_t)\|} \right\}^{-c\beta(y;x_t)}$$

*if $\mu_t - \mu(x_t)$ and $y_t - \mu(x_t)$ are linearly independent,*

$$q(y; \mu_t, x_t) = \left\{ \frac{|[y - \mu(x_t)]'(y - \mu_t)|}{\|y - \mu(x_t)\|} \right\}^{-c\beta(y;x_t)\mathrm{sgn}\{[y-\mu(x_t)]'(y-\mu_t)\}}$$

*otherwise, and*

$$\beta(y; x_t) = \frac{[y - \mu(x_t)]'\Sigma(x_t)^{-1}[y - \mu(x_t)]}{\|y - \mu(x_t)\| \, \|\Sigma(x_t)^{-1}[y - \mu(x_t)]\|}.$$

*The function $\beta(y; x_t)$ takes values in the unit interval $[0, 1]$; it is identically equal to unity if $\Sigma(x_t)$ is proportional to the identity matrix.*

**Remark 1** *In the univariate case, the function $q$ reduces to $q\,[y; \mu_t, \mu(x_t)] = |y - \mu_t|^{-c}$ if $y \notin (\mu_t; \mu(x_t))$, and $|y - \mu_t|^c$ if $y \in (\mu_t; \mu(x_t))$. The robust filter provided by Proposition 4 coincides with the univariate solution obtained in Section 3.2.*

**Construction of the Robustified Gaussian**

<u>Step 1 (Critical Roots)</u>: Compute the real roots $s_1 < ... < s_J$ in the unit interval (see Appendix) of the equation

$$s^2 \left\|\Sigma(x_t)^{-1}\tilde{y}_t\right\|^2 \left(s^2 \|\tilde{y}_t\|^2 - 2s\tilde{\mu}_t'\tilde{y}_t + \|\tilde{\mu}_t\|^2\right) - c^2 = 0\,, \tag{3.11}$$

where $\tilde{y}_t = y_t - \mu(x_t)$, $\tilde{\mu}_t = \mu_t - \mu(x_t)$.

<u>Step 2 (Critical Threshpoints)</u>: Calculate $y_t^{(j)} = \mu(x_t) + s_j[y_t - \mu(x_t)]$ for $s_0 = 0 < s_1 < ... < s_J < s_{J+1} = 1$.

<u>Step 3 (Robustified observation density)</u>: Compute $\tilde{f}(y_t|x_t, Y_{t-1})$ equal to

$$f_{\mathcal{N}}[y_t^{(1)}; \mu(x_t), \Sigma(x_t)] \prod_{\substack{j=1 \\ j \text{ even}}}^{J} \frac{f_{\mathcal{N}}[y_t^{(j+1)}; \mu(x_t), \Sigma(x_t)]}{f_{\mathcal{N}}[y_t^{(j)}; \mu(x_t), \Sigma(x_t)]} \prod_{\substack{j=1 \\ j \text{ odd}}}^{J} \frac{q[y_t^{(j+1)}; \mu_t, x_t]}{q[y_t^{(j)}; \mu_t, x_t]}$$

**Remark 2** *If $\mu(x_t) = \mu_t$ and $\Sigma(x_t) = \sigma^2(x_t)I_{p\times p}$, the nonnormalized robustified density coincides with the Gaussian if $\|y - \mu_t\| \leq \sqrt{c}\sigma(x_t)$, and is equal to*

$$\tilde{f}(y|x_t, Y_{t-1}) = \frac{1}{(2\pi)^{p/2} e^{c/2} \sigma(x_t)^p} \left(\frac{\|y - \mu\|}{\sigma(x_t)\sqrt{c}}\right)^{-c} \tag{3.12}$$

*if $\|y - \mu_t\| > \sqrt{c}\sigma(x_t)$.*

## 3.4 Choosing the Tuning Constant

The following proposition can be used to select the tuning constant.

**Proposition 5 (Efficiency of the robustified Gaussian)** *If the observation density $f(y_t|x_t, Y_{t-1})$ is a spherical multivariate Gaussian $\mathcal{N}[\mu(x_t), \sigma^2(x_t)I_{p\times p}]$ with mean $\mu(x_t) = \mu_t$ for every $x_t$, then the normalizing constant is independent of the state:*

*$B_t(x_t) = B$ for all $x_t$. The constant $B$ is if $p = 1$ :*

$$B = \left[2\Phi(\sqrt{c}) - 1 + \frac{2\sqrt{c}}{c-1}\phi(\sqrt{c})\right]^{-1},$$

*if $p$ is odd and $p \geq 3$ :*

$$\begin{aligned} B =& \Big\{1 + \frac{e^{-c/2}}{2^{p/2-1}\Gamma(p/2)}\Big[\frac{c^{p/2}}{c-p} - \frac{(p-2)!\sqrt{2c}}{2^{p/2-2}((p-3)/2)!}\sum_{i=0}^{(p-3)/2}\frac{2^i(i+1)!}{(2i+2)!}c^i\Big] \\ & - \frac{2\sqrt{\pi}}{2^{p-2}\Gamma(p/2)}\frac{(p-2)!}{((p-3)/2)!}[1 - \Phi(\sqrt{c})]\Big\}^{-1}, \end{aligned}$$

*and if $p$ is even:*

$$B = \left\{1 + \frac{e^{-c/2}}{2^{p/2-1}\Gamma(p/2)}\left[\frac{c^{p/2}}{c-p} - \left(\frac{p}{2}-1\right)!\,2^{p/2-1}\sum_{i=0}^{p/2-1}\frac{1}{i!}\left(\frac{c}{2}\right)^i\right]\right\}^{-1}.$$

*In the above equations, the tuning constant $c$ is strictly larger than $p$, $\Gamma$ is Euler's Gamma function, and $\phi(\cdot)$ and $\Phi(\cdot)$ respectively denote the density and cumulative distribution function of a standard normal. Furthermore, the backward Kullback-Leibler divergence in (3.5) satisfies*

$$KL_t^{eff} \leq -\log(B)\,.$$

For a given deviation $\alpha$, the analyst can choose the tuning constant $c$ such that:

$$-\log(B) = \alpha\,.$$

In Figure 4 we plot $-\log(B)$ as a function of $c$ for spherical Gaussians of dimensions 1 to 4 centered around $\mu_t$. Table 1 reports the tuning constants corresponding to

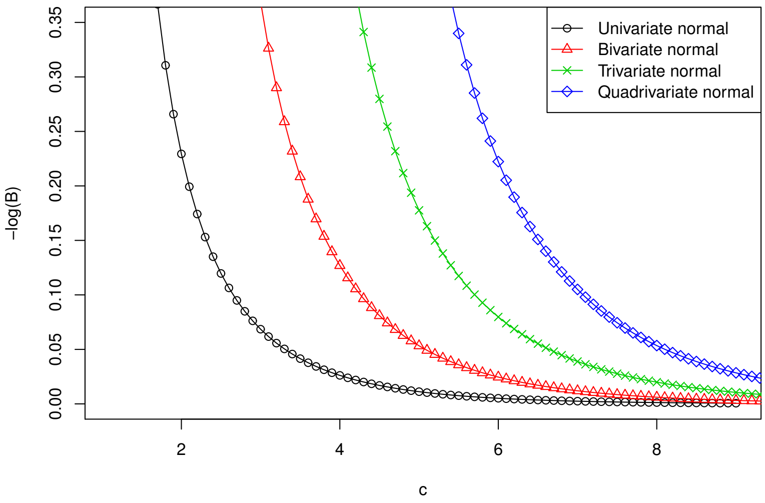

Figure 4: Link between the tuning constant $c$ and the normalizing constant $B$.

$\alpha = 0.05$ and 0.01.

### 3.5 An Alternative Robustification via Student $t$

In the special case where the observation density $f(y_t|x_t, Y_{t-1})$ is a spherical Gaussian $\mathcal{N}[\mu_t, \sigma^2(x_t)I_{p\times p}]$ centered around $\mu_t$, we can construct an alternative robustified filter by using a noncentered multivariate Student $t$ distribution with $\nu$ degrees of freedom:

$$\hat{f}_S(y_t|x_t, Y_{t-1}) = \frac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)\,(\nu+p)^{p/2}\,\pi^{p/2}\,\sigma(x_t)^p}\left[1 + \frac{\|y_t - \mu_t\|^2}{(\nu+p)\sigma^2(x_t)}\right]^{-(\nu+p)/2} . \tag{3.13}$$

Since

$$\left\|\frac{\partial \log \hat{f}_S(y_t|x_t, Y_{t-1})}{\partial y_t}\right\| \|y_t - \mu_t\| = \frac{(\nu+p)\|y_t - \mu_t\|^2}{(\nu+p)\sigma^2(x_t) + \|y_t - \mu_t\|^2} < \nu + p, \tag{3.14}$$

Table 1: Selection of the tuning constant $c$

| $p$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
|---|---|---|
| 1 | 3.3091 | 5.1413 |
| 2 | 5.0786 | 7.2646 |
| 3 | 6.6405 | 9.0844 |
| 4 | 8.1043 | 10.7618 |

the function $\hat{f}_S$ satisfies robustness condition (2.3) for $c = \nu + p$. In Section 4.3, we will investigate in the context of a stochastic volatility model the behavior of a robust filter using a Student distribution $\hat{f}_S$ with $\nu = c - p$ degrees of freedom.

## 3.6 Implementation

Robust filtering can be applied to a variety of models.

### 3.6.1 Finite State Space

When the state space is finite and contains $d$ possible elements $m^1, \ldots, m^d$, the Bayesian filter is available analytically (Hamilton, 1989; Lindgren, 1978):

$$\lambda_t = \frac{\omega_t(y_t) \odot (\lambda_{t-1} A)}{[\omega_t(y_t) \odot (\lambda_{t-1} A)]\mathbf{1}'}\,, \tag{3.15}$$

where $\lambda_t = [\mathbb{P}(x_t = m^j | Y_t)]_{1 \leq j \leq d}$ denotes the row vector of filtered probabilities, $\omega_t(y_t) = [f(y_t | x_t = m^j, Y_{t-1})]_{1 \leq j \leq d}$ the row vector of observation densities, $A = (a_{i,j})_{1 \leq i,j \leq d}$ the transition matrix with elements $a_{i,j} = \mathbb{P}(x_{t+1} = m^j | x_t = m^i)$, $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^d$, and $\odot$ denotes the Hadamard product: $x \odot y = (x_1 y_1, \ldots, x_d y_d)$ for all $x, y \in \mathbb{R}^d$. In applications, the initial filter $\lambda_0$ is usually set equal to the ergodic distribution. The log-likelihood of the dataset $Y_T$ is $\mathcal{L}(Y_T) = \sum_{t=1}^{T} \log[\omega_t(y_t)(\lambda_{t-1}A)']$.

We obtain a robust filter $\hat{\lambda}_t$ by replacing the noncontaminated observation density with its robust version (3.2) in the recursion (3.15):

$$\hat{\lambda}_t = \frac{\hat{\omega}_t(y_t) \odot (\hat{\lambda}_{t-1}A)}{[\hat{\omega}_t(y_t) \odot (\hat{\lambda}_{t-1}A)]\mathbf{1'}}\,, \tag{3.16}$$

where $\hat{\omega}_t(y_t) = [\tilde{f}(y_t|x_t = m^j, Y_{t-1})]_{1\leq j\leq d}$ is the row vector of nonnormalized robustified observation densities. The *robustified log-likelihood* is then

$$\hat{\mathcal{L}}(Y_T) = \sum_{t=1}^{T} \log[\hat{\omega}_t(y_t)(\hat{\lambda}_{t-1}A)']. \tag{3.17}$$

Note that the use of nonnormalized robustified densities in the definition of $\hat{\omega}_t(y_t)$ is driven by computational convenience. Indeed, consider the alternative filter based on the normalized densities $\hat{f}(\cdot|x_t, Y_{t-1})$. By (3.17), the alternative filter coincides with the former filter $\hat{\lambda}_t$ when the normalizing constant $B_t(x_t)$ is state-invariant, as is the case with the Gaussians $\mathcal{N}[\mu_t, \sigma^2(x_t)\, I_{p\times p}]$ investigated in Proposition 5. In general, however, the alternative filter requires the computation of the normalizing constants $B_t(x_t)$ for every state in every period, which is numerically expensive. For this reason, we do not pursue this alternative definition and use only the robustified filter $\hat{\lambda}_t$ in the rest of the paper.

### 3.6.2 Particle Filtering

In more complex environments, Bayesian inference can proceed via particle filters, defined as a set $\{x_t^{(n)}\}_{n=1}^{N}$ that discretely approximates the filtering distribution $\lambda(x_t|Y_t)$. More specifically, we consider bootstrap filters such as those by Gordon, Salmond and Smith (1993), which are defined by the following recursive three-step procedure.

1. Particles $\{\tilde{x}_t^{(n)}\}_{n=1}^N$ are generated from $\{x_{t-1}^{(n)}\}_{n=1}^N$ using $\rho(\cdot|x_{t-1}^{(n)})$.
2. Each particle $\tilde{x}_t^{(n)}$, $n = 1, \ldots, N$, is given an importance weight $w_t^{(n)} = f(y_t|\tilde{x}_t^{(n)}, Y_{t-1})$.
3. A new set of particles $\{x_t^{(n)}\}_{n=1}^N$ is drawn from a multinomial distribution with support $\{\tilde{x}_t^{(n)}\}_{n=1}^N$ and associated probabilities $\{\pi_t^{(n)}\}_{n=1}^N$, where $\pi_t^{(n)} = w_t^{(n)}/\sum_{n'} w_t^{(n')}$.

The one-step likelihood $f(y_t|Y_{t-1})$ is consistently estimated by $N^{-1}\sum_{n=1}^N w_t^{(n)}$ and $\mu_t = \mathbb{E}(y_t|Y_{t-1})$ by the sample mean of $\{\mu(\tilde{x}_t^{(n)})\}_{n=1}^N$. We define the robust particle filter as a modified version of this algorithm, where in Step 2 the importance weights $w_t^{(n)}$ are replaced by their robustified version $\hat{w}_t^{(n)} = \tilde{f}(y_t|\tilde{x}_t^{(n)}, Y_{t-1})$.

The robust filter naturally solves the degeneracy problems that plague classical particle filters, as Figure 2 illustrates. If $y_t$ is an outlier in the right tail of the figure, the classical filter assigns quickly declining importance weights $w_t^{(n)} = f(y_t|x_t^{(n)}, Y_{t-1})$ to particles $\tilde{x}_t^{(n)}$ associated with a high mean; in Step 3 of the procedure, the filter can therefore "collapse" into the particle with the largest weight. By contrast, the robust filter assigns more evenly distributed weights to states $\hat{w}_t^{(n)}$ with a positive mean, so that a wide range of particles are drawn in Step 3. In Section 4.3, we will verify by Monte Carlo simulation the validity of this intuition.

## 3.7 Piecewise Differentiability and Optimality

We have so far focused on filters that are continuously differentiable on the entire observation space. More generally, one can consider piecewise differentiable filters.

Assume for simplicity that $y_t \in \mathbb{R}$, and let $\mathcal{PC}^1(\mathbb{R})$ denote the space of functions that are continuous everywhere and piecewise continuously differentiable on the real line. If the observation density $f(\cdot|x_t, Y_{t-1})$ belongs to $\mathcal{PC}^1(\mathbb{R})$ for every $x_t$ and $Y_{t-1}$, then one can easily rewrite Definition 2.2 (Impact function) and Definition 2 (Robustness) with left and right derivatives. Under these definitions, the filtering distribution $\lambda$ is robust if the observation density satisfies the conditions

$$|(y_t - \mu_t)\partial_- \log f(y_t|x_t, Y_{t-1})| \leq c \tag{3.18}$$

$$|(y_t - \mu_t)\partial_+ \log f(y_t|x_t, Y_{t-1})| \leq c \tag{3.19}$$

for every $y_t$, $x_t$, and $Y_{t-1}$.

These extensions permit the construction of new robustified filters. For instance if $x_t$ is fixed, we can consider two threshpoints $\underline{y}^*$ and $\bar{y}^*$ such that $\underline{y}^* < \mu_t < \bar{y}^*$, and define the function:

$$f^*(y|x_t, Y_{t-1}) = \begin{cases} C^*_{1,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t < \underline{y}^*, \\ B^*_t(x_t)f(y|x_t, Y_{t-1}) & \text{if } y_t \in [\underline{y}^*, \bar{y}^*], \\ C^*_{2,t}(x_t)|y_t - \mu_t|^{-c} & \text{if } y_t > \bar{y}^*, \end{cases} \tag{3.20}$$

where $C^*_{1,t}(x_t) = B^*_t(x_t)f(\underline{y}^*|x_t, Y_{t-1})|\underline{y}^* - \mu_t|^c$ and $C^*_{2,t}(x_t) = B^*_t(x_t)f(\bar{y}^*|x_t, Y_{t-1})|\bar{y}^* - \mu_t|^c$. The function $f^*$ has a unit integral if the normalizing constant $B^*_t(x_t)$ is selected as follows:

$$\begin{aligned} [B^*_t(x_t)]^{-1} &= F(\bar{y}^*|x_t, Y_{t-1}) - F(\underline{y}^*|x_t, Y_{t-1}) \\ &\quad + (c-1)^{-1}\left[(\mu_t - \underline{y}^*)\, f(\underline{y}^*|x_t, Y_{t-1}) + (\bar{y}^* - \mu_t)\, f(\bar{y}^*|x_t, Y_{t-1})\right], \end{aligned} \tag{3.21}$$

where $F(y|x_t, Y_{t-1}) = \int_{-\infty}^{y} f(z|x_t, Y_{t-1})dz$ denotes the cumulative distribution function of the observation density. The class of functions defined by (3.20) and (3.21) generalizes the normalized robustified density constructed in earlier sections. When the threshpoints $\underline{y}^*$ and $\bar{y}^*$ are appropriately selected, the function $f^*$ is optimal in the Kullback-Leibler sense.

**Proposition 6** *We consider that the observation density satisfies Assumptions 2 and 4 and that the tuning constant exceeds unity:* $c > 1$*. If the transition points* $\underline{y}^*$ *and* $\bar{y}^*$ *solve the equations*

$$\frac{f(\underline{y}^*|x_t, Y_{t-1})(\mu_t - \underline{y}^*)}{F(\underline{y}^*|x_t, Y_{t-1})} = c - 1, \tag{3.22}$$

$$\frac{f(\bar{y}^*|x_t, Y_{t-1})(\bar{y}^* - \mu_t)}{1 - F(\bar{y}^*|x_t, Y_{t-1})} = c - 1, \tag{3.23}$$

*then the function* $f^*(y_t|x_t, Y_{t-1})$ *defined by (3.20) and (3.21) is the unique solution to the program:*

$$\min_{h(\cdot|x_t, Y_{t-1}) \in \mathcal{PC}^1(\mathbb{R})} KL\left[f(y_t|x_t, Y_{t-1}); h(y_t|x_t, Y_{t-1})\right] \tag{3.24}$$

*subject to* $h(y_t|x_t, Y_{t-1}) > 0$, $\int_{\mathbb{R}} h(y_t|x_t, Y_{t-1})dy_t = 1$, *and the robustness constraints (3.18)–(3.19) for every* $y_t \in \mathbb{R}$.

*More generally for any* $m > 0$, *the function* $m\, f^*(y_t|x_t, Y_{t-1})$ *is the unique solution to (3.24) subject to*

$$\int_{\mathbb{R}} h(y_t|x_t, Y_{t-1})dy_t = m,$$

*and the maintained positivity and robustness constraints.*

The optimum $f^*(\cdot|x_t, Y_{t-1})$ is the probability density function which is the closest to the model's $f(\cdot|x_t, Y_{t-1})$ according to the Kullback-Leibler divergence (efficiency measure), subject to a bound of the impact function (robustness measure). Moreover, the optimum is piecewise differentiable in the observation vector $y$ and exhibits corners at $\underline{y}$ and $\bar{y}$. It is therefore a "broken extremum" in the terminology of the calculus of variations.

Related optimization problems also have broken extrema. Thus in the class of robust functions with the same mass as the nonnormalized filter $\tilde{f}(\cdot|x_t, Y_{t-1})$ in (3.2), the rescaled function $[B_t(x_t)]^{-1}\ f^*(\cdot|x_t, Y_{t-1})$ minimizes the Kullback Leibler divergence to $f(\cdot|x_t, Y_{t-1})$. One can also consider criteria other than (3.24) and verify that broken extrema with two corners (not shown here) minimize the $L^1$ or $L^2$ distance to the model's observation density.

These results are in the spirit of the optimality properties previously obtained for robust estimators and tests (Hampel et al., 1986) and for the robust Kalman filter (Ruckdeschel, 2010a), as will be discussed in Section 4.2. In unreported simulations, we have observed that the optimum $f^*(\cdot|x_t, Y_{t-1})$ does not substantially improve on the robustified density $\tilde{f}(\cdot|x_t, Y_{t-1})$ in terms of filtering efficiency. Because the optimum $f^*(\cdot|x_t, Y_{t-1})$ contains corners, however, it generates likelihoods that vary irregularly with the structural parameters of the model, which induces large losses in estimation efficiency. The filter $f^*(\cdot|x_t, Y_{t-1})$ is thus not operational for likelihood-based estimation or model selection. For this reason, we henceforth focus on the general-purpose methodology based on the nonnormalized robustified density $\tilde{f}(\cdot|x_t, Y_{t-1})$, which is fully described in Section 3.6.2.

# 4 Applications

We now present several applications of robust filtering. Section 4.1 considers filtering and model selection in finite-state space models, for which the robust filter is available in closed form. Section 4.2 focuses on linear Gaussian state-space models and investigates how the robust particle filter performs relative to the original and an earlier robust version of the Kalman filter. Section 4.3 applies robust filtering to a stochastic volatility model.

## 4.1 Finite State-Space Model

We begin our investigation with finite state-space models of the type described in Section 3.6.1.

### 4.1.1 A Univariate Multifractal Model

We consider a Markov-Switching Multifractal (MSM) model, as defined in Calvet and Fisher (2001) and Calvet and Fisher (2008). That is, the state $x_t$ has $\overline{k}$ components that can each take two values, $m_0$ or $2 - m_0$. Consequently, the state $x_t \in \{m_0, 2 - m_0\}^{\overline{k}}$ takes $d = 2^{\overline{k}}$ values $m^1, \ldots, m^d$. The transition matrix $A$ defined in Section 3.6.1 has elements

$$a_{i,j} = \prod_{l=1}^{\overline{k}} \left[ \left(1 - \frac{\gamma_l}{2}\right) \mathbb{1}_{m_l^i = m_l^j} + \frac{\gamma_l}{2} \mathbb{1}_{m_l^i \neq m_l^j} \right] \tag{4.1}$$

Table 2: Model selection, robust filter with $\alpha = 0.01$

| | Number of times model with $\overline{k}$ components is selected | | | | | | | | Prop. of correct sel. |
|---|---|---|---|---|---|---|---|---|---|
| | $\overline{k}=1$ | $\overline{k}=2$ | $\overline{k}=3$ | $\overline{k}=4$ | $\cdots$ | $\overline{k}=8$ | $\overline{k}=9$ | $\overline{k}=10$ | (in %) |
| $\mathcal{L}$, no cont. | 0 | 0 | 100 | 0 | $\cdots$ | 0 | 0 | 0 | 100 |
| Rob. $\mathcal{L}$, no cont. | 0 | 0 | 100 | 0 | $\cdots$ | 0 | 0 | 0 | 100 |
| PF, no cont. | 0 | 0 | 100 | 0 | $\cdots$ | 0 | 0 | 0 | 100 |
| RobPF, no cont. | 0 | 0 | 100 | 0 | $\cdots$ | 0 | 0 | 0 | 100 |
| $\mathcal{L}$, 5% cont. | 0 | 0 | 0 | 0 | $\cdots$ | 0 | 3 | 97 | 0 |
| Rob. $\mathcal{L}$, 5% cont. | 0 | 0 | 100 | 0 | $\cdots$ | 0 | 0 | 0 | 100 |
| PF, 5% cont. | 0 | 0 | 0 | 0 | $\cdots$ | 0 | 4 | 96 | 0 |
| RobPF, 5% cont. | 0 | 0 | 100 | 0 | $\cdots$ | 0 | 0 | 0 | 100 |

with $\gamma_l = 1 - (1-\gamma_1)^{b^{l-1}}$ for all $l = 1, \ldots, \overline{k}$. The observations are given by

$$y_t = \sigma(x_t)\epsilon_t,$$
$$\sigma(x_t) = \overline{\sigma}\left(\Pi_{i=1}^{\overline{k}} x_{t,i}\right)^{1/2},$$

where the random variables $\epsilon_t$ are independent standard normals. MSM is specified by the parameter vector $\theta = (m_0, \gamma_1, b, \overline{\sigma})$. In all simulations, we set $m_0 = 1.5$, $\gamma_1 = 0.0005$, $b = 2$ and $\overline{\sigma} = 1$.

We first generate an uncontaminated dataset $Y_T$ of size $T = 10^3$ from an MSM with $\overline{k} = 6$ components. We then contaminate the dataset with replacement outliers:

$$y_t^{cont} = \eta\, y_t, \tag{4.2}$$

which replace $y_t$ with probability 5% at every date $t$. We set $\eta = 4$ in simulations. In the notation of Maronna, Martin and Yohai (2006) p. 253, the statistician observes $(1 - z_t)\, y_t + z_t\, y_t^{cont}$, where $y_t^{cont}$ is the replacement process and the $z_t's$ are independent Bernoulli variates such that $\mathbb{P}(z_t = 1) = 0.05$. Note that the perturbation is unbounded and therefore more challenging to address than the bounded contaminations considered in Sections 2 and 3.

Since the state-space is finite and the observation density Gaussian, we can implement the robust filter defined in Section 3.6.1 using the robustified Gaussian observation density defined in Section 3.2. The tuning constant is $c = 5.1413$ (see Table 1 with $\alpha = 0.01$). Figure 1 reports at each date $t$ the standard and robust filters as measured by the $KL_t$ divergence in (1.3) under uncontaminated (left panel) and 5% contaminated (right panel) data. Figure 1 illustrates that the robust filter is much less sensitive to contamination than the standard Bayesian filter. We now turn to likelihood-based model selection. To compare the model selection accuracy of the various filtering methods, we generate 100 samples from a univariate MSM with $\overline{k} = 3$ components. We estimate the likelihood functions for uncontaminated and 5%-contaminated cases and various $\overline{k} = 1, \ldots, 10$ using four methods: standard and robust Bayesian updating, standard and robust particle filter with $N = 10^5$ highlighted in Section 3.6.2.

In Table 2, we report for each filtering method the number of times a particular specification $\overline{k}$ achieves the highest likelihood and is therefore selected. Under no contamination, all four methods consistently select the right model specification $\overline{k} = 3$. Under 5% contamination, standard filters are highly inaccurate and select models with the largest number of volatility components considered ($\overline{k} = 10$). The two robust filters, however, consistently select the correct model specification.

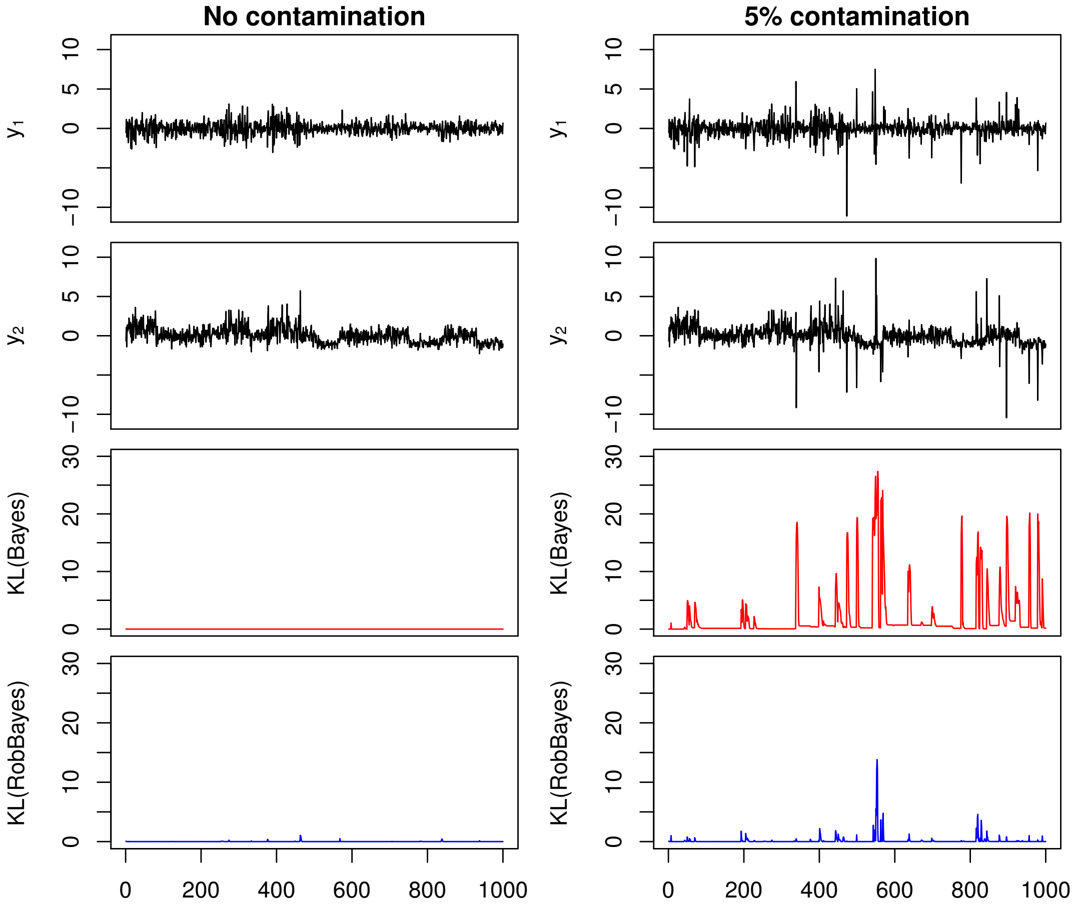

Figure 5: Accuracy of standard and robust Bayesian filters in a bivariate MSM without contamination (left panel) and with 5% contamination (right panel).

### 4.1.2 Tracking the State of a Bivariate Multifractal Model

We consider the bivariate MSM model defined in Calvet, Fisher and Thompson (2006):

$$y_t = \mu(x_t) + \Sigma(x_t)\epsilon_t \, , \tag{4.3}$$

where $x_t$ is defined in the previous section,

$$\mu(x_t) = \begin{pmatrix} 0 \\ \sum_{i=1}^{\overline{k}}(x_{t,i} - 1) \end{pmatrix}, \qquad \Sigma(x_t) = \left(\Pi_{i=1}^{\overline{k}} x_{t,i}\right)^{1/2} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \tag{4.4}$$

and $\epsilon_t$ are independent bivariate normal variables with expectation 0, variances 1, and correlation $\tilde{\rho}$. In all simulations, the parameter values are $\gamma_1 = 0.0005$, $m_0 = 1.5$, $b = 2$, $\tilde{\rho} = 0.5$, and $\sigma_1 = \sigma_2 = 1$.

Consider an uncontaminated process $Y_t$ drawn from (4.3)–(4.1) and a 5%-RO contaminated process:

$$y_t^{cont} = y_t + \xi_t \tag{4.5}$$

where $\xi_t \sim \mathcal{N}(0, \eta^2 I_{2\times 2})$. We use $\eta = 4$ in simulations.

Figure 5 illustrates the accuracy measure $KL_t$ in (1.3) for uncontaminated (left panel) and 5%-contaminated (right panel) bivariate MSM samples of size $T = 10^3$ with $\overline{k} = 6$. The robust filter in Section 3.6.1 is applied here with a robustified bivariate Gaussian with tuning constant $c = 7.2646$ (see Table 1 with $\alpha = 0.01$). Consistent with the univariate results of the previous section, the robust filter achieves considerable gains in accuracy compared to the standard Bayesian filter under contamination.

## 4.2 Linear Gaussian State-Space Model

We consider the linear Gaussian model

$$y_t = Hx_t + u_t \,,$$
$$x_t = Fx_{t-1} + w_{t-1} \,,$$

where $y_t \in \mathbb{R}^p$, $x_t \in \mathbb{R}^m$, $u_t \sim \mathcal{N}(0, R)$, $w_t \sim \mathcal{N}(0, Q)$, and $\mathbb{E}(u_t\, w_t') = 0$. In simulations, we set $p = m = 2$,

$$F = \begin{pmatrix} 0.9 & 0 \\ 0 & 0.9 \end{pmatrix}, \; H = \begin{pmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{pmatrix},$$

and $R = Q = I_{2\times 2}$.

*Kalman Filter ("KF").* The state distribution $x_t|Y_t \sim \mathcal{N}(\hat{x}_t, P_t)$ can be estimated via Kalman's algorithm:

$$\overline{x}_t = F\hat{x}_{t-1} \,, \tag{4.6}$$

$$M_t = FP_{t-1}F' + Q \,, \tag{4.7}$$

$$\Sigma_t = HM_tH' + R \,, \tag{4.8}$$

$$K_t = M_tH'\Sigma_t^{-1} \,, \tag{4.9}$$

$$\hat{x}_t = \overline{x}_t + K_t\epsilon_t \,, \tag{4.10}$$

$$\epsilon_t = y_t - H\overline{x}_t \,, \tag{4.11}$$

$$P_t = M_t - K_tHM_t' \,. \tag{4.12}$$

We initialize the filter with $\hat{x}_0 = 0$ and the stationary value $P_0 = (1 - 0.9^2)^{-1}I_{2\times 2}$.

*Robust Kalman Filter ("RobKF").* A robust version of the Kalman filter can be obtained by "huberizing" the prediction error $\epsilon_t$ in (4.10):

$$\hat{x}_t = \overline{x}_t + K_t \epsilon_t \min\left(1, \frac{\kappa}{\|K_t \epsilon_t\|}\right) . \tag{4.13}$$

This is a natural operation, which has been used in many different models in order to obtain robust estimators and tests. Indeed the Huber function bounds the unlimited influence of a single observation $y_t$ (through $\epsilon_t$) on the filter. Moreover, as shown by Theorem 3.2 in Ruckdeschel (2010a), the robust Kalman filter obtained by replacing (4.10) by (4.13) has two optimality properties. First, it is minimax in the sense that it minimizes the worst mean squared error over a neighborhood of the underlying model distribution. Secondly, it minimizes the mean squared error, subject to a bound of the bias in the neighborhood.

*Filter Comparison.* We now compare the accuracy of KF, RobKF, the bootstrap filter ("PF") of Gordon, Salmond and Smith (1993) and the robust particle filter defined in Section 3 ("RobPF"). We consider contaminations defined in (4.5) and $\eta = 4$. Note that an exact Kalman filter is available when the contamination (4.5) is known; it consists of the algorithm (4.6)–(4.12) except that (4.8) is replaced by $\Sigma_t^{cont} = \Sigma_t + \eta^2 I_{2\times 2}$ when an outlier occurs. We denote the associated mean and variance processes by $\{\hat{x}_t^{cont}\}$ and $\{P_t^{cont}\}$.

The accuracy measure $KL_t$ in (1.3) is here the backward Kullback-Leibler divergence between the Gaussian distributions $\mathcal{N}(\hat{x}_t, P_t)$ and $\mathcal{N}(\hat{x}_t^{cont}, P_t^{cont})$:

$$KL_t = \frac{1}{2}\left\{\mathrm{tr}[(P_t^{cont})^{-1} P_t] + (\hat{x}_t^{cont} - \hat{x}_t)'(P_t^{cont})^{-1}(\hat{x}_t^{cont} - \hat{x}_t) - \log\left(\frac{\det P_t}{\det P_t^{cont}}\right) - p\right\} .$$

Figure 6 illustrates the divergence computed every period with the KF, RobKF, PF,

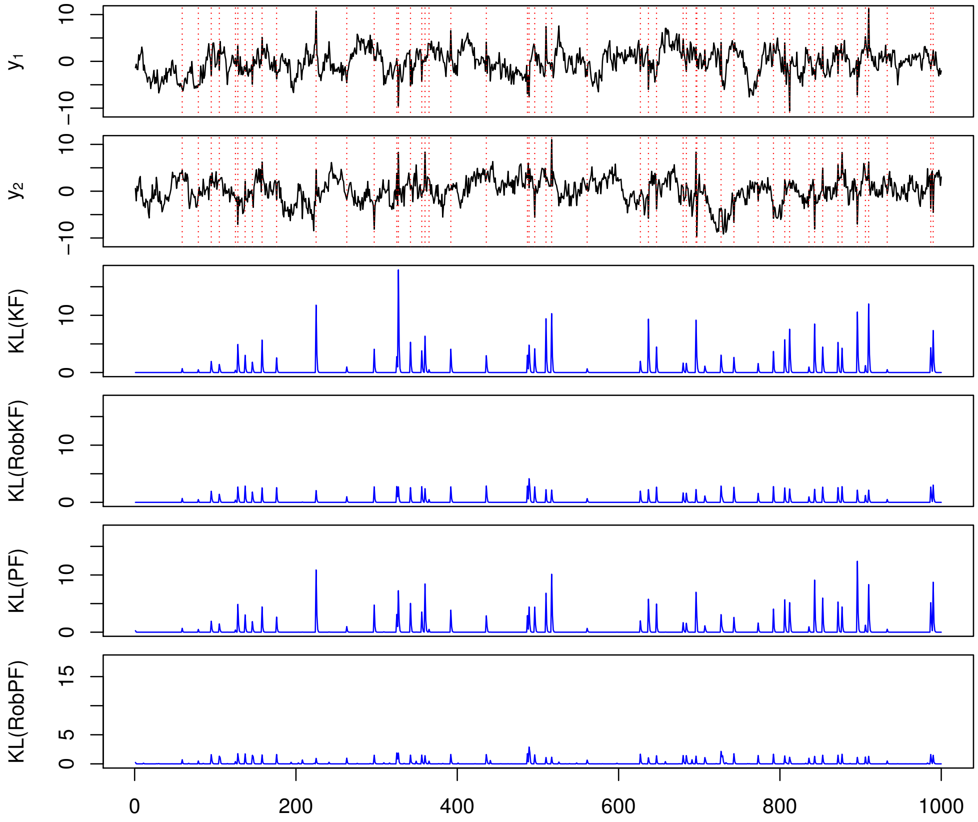

Figure 6: Kullback-Leibler divergences between normal distributions with filtered estimates of $\mu(x_t)$ and $\Sigma(x_t)$, bivariate case with 5% contamination.

and RobPF filters. RobPF is implemented with the tuning constant $c = 7.2646$, (see Table 1), while the RobKF method (4.13) is applied with $\kappa = 1.345/\sqrt{1 - 0.9^2} = 3.0856$ based on the stationary variance $P_0$. The sample size is $T = 10^3$ and the particle filter size $N = 10^4$. The maximum value of $\{KL_t\}_{t=1,\dots,T}$ is 17.95 for KF, 4.14 for RobKF, 12.40 for PF and 2.90 for RobPF. (In the noncontaminated case, the corresponding maximum values are 0, 0.51, 0.26 and 0.69, respectively.)

We next generate 100 samples with $\eta = 0, 2, 4$ and for each sample and filter compute the average Kullback-Leibler divergence $KL^{\text{eff},a} = T^{-1} \sum_{t=1}^{T} KL_t$. The ag-

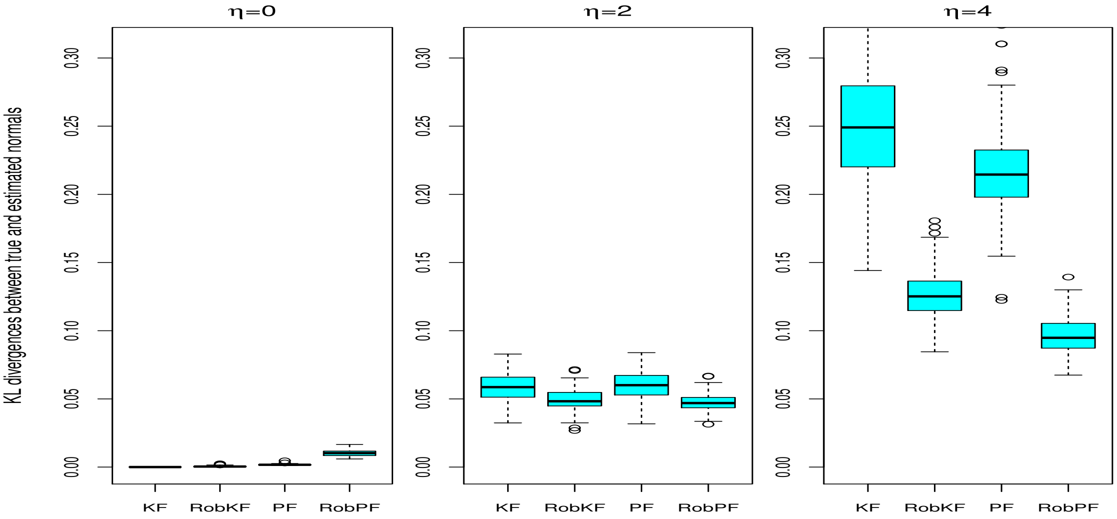

Figure 7: Boxplots of aggregate Kullback-Leibler divergences between normal distributions with filtered estimates of $\mu(x_t)$ and $\Sigma(x_t)$, bivariate case with 5% contamination $\mathcal{N}(0, \eta^2 I_{2\times 2})$. (Notice that $KL \equiv 0$ for the Kalman filter for $\eta = 0$.)

gregate $KL^{\text{eff},a}$ measures are reported in the boxplots of Figure 7. Both figures show the extreme sensitivity of the nonrobust methods (KF and PF) to a small amount of contamination. Their robust versions lose very little accuracy under no contamination, but provide large gains under contamination. Finally, notice that although the robust particle filter is not tuned for this particular application, its accuracy is similar to that of the robust Kalman filter which is specifically designed for Gaussian linear models.

## 4.3 Stochastic Volatility

Stochastic volatility models play an important role in finance and particle filters are often used to estimate them; see e.g. Chib, Nardari and Shephard (2006) and

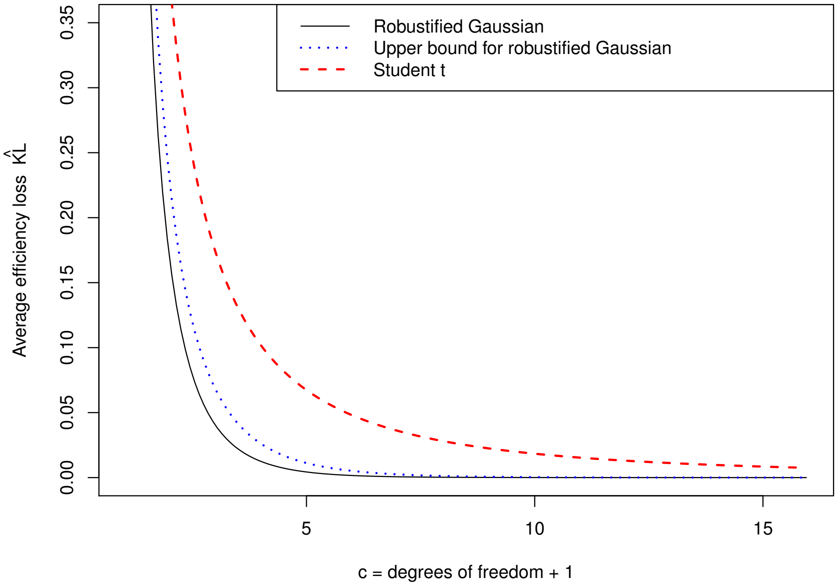

Figure 8: Efficiency measures in an uncontaminated SV simulation, using robustified Gaussian and Student $t$ particle filters with $c = \nu + 1$.

Johannes and Polson (2009). We consider the popular specification

$$\begin{aligned} y_t &= e^{x_t/2}\epsilon_t\,, \\ x_t &= a + b\,x_{t-1} + \sigma u_t\,, \end{aligned} \tag{4.14}$$

with $\epsilon_t$ and $u_t$ independent standard normal variables. In simulations, we let $a = -0.005$, $b = 0.99$ and $\sigma = 0.1$.

Since $y_t|x_t \sim \mathcal{N}(0, e^{x_t})$, the robustified density (3.8) and the Student $t$ variant (3.13) can both be applied. By (3.14), the robustified Gaussian with tuning constant $c$ and the Student $t$ with $\nu = c - 1$ provide similar robustness levels, in the sense that both functions satisfy condition (2.3) with an upper-bound equal to $c$.

We now compare the efficiency costs of the two robust methods for a variety of tuning constants $c > 1$. We generate a dataset of size $T = 10^3$ from the uncontaminated model (4.14), which is illustrated in the top left panel of Figure 9. For a given robust filter, we can measure the average efficiency cost by

$$\widehat{KL}^{\text{eff},a} = \frac{1}{T}\sum_{t=1}^{T}\widehat{KL}_t^{\text{eff}}, \tag{4.15}$$

where

- $\widehat{KL}_t^{\text{eff}} = (NK)^{-1}\sum_{m=1}^{N}\sum_{k=1}^{K}\left\{\log\left[\frac{\sum_{n=1}^{N} f(\tilde{y}_t^{(k)}|\tilde{x}_t^{(n)},Y_{t-1})}{\sum_{n=1}^{N}\hat{f}(\tilde{y}_t^{(k)}|\tilde{x}_t^{(n)},Y_{t-1})}\right]\right\}$ is a particle filter estimator of the divergence $KL_t^{\text{eff}}$;
- $\{\tilde{x}_t^{(n)}\}_{n=1}^{N}$ are the Step 1 particles of the standard bootstrap filter;
- for each $m = 1,\ldots,N$, the values $\{\tilde{y}_t^{(k)}\}_{k=1,\ldots,K}$ are sampled from $\mathcal{N}(0, e^{\tilde{x}_t^{(m)}})$.

Figure 8 plots $\widehat{KL}^{\text{eff},a}$ for the two robust particle filters for various $c$ and $K = N = 200$. For convenience, Figure 8 also reports the upper-bound for $KL_t^{\text{eff}}$ considered in Proposition 5 and Figure 4. We see that for any robustness level, the robustified Gaussian particle filter always results in a smaller efficiency cost than its Student $t$ equivalent. Note also that the upper-bound derived in Proposition 5 is a very accurate approximation of the actual efficiency cost of the robustified Gaussian particle filter.

For a given level of desired efficiency cost $\alpha$, we can choose the corresponding constants $c$ and $\nu$. For instance, we reach $\widehat{KL}^{\text{eff},a} = 0.05$ with a robustified Gaussian with $c = 2.8$ and a Student $t$ with $\nu = 4.9$ degrees of freedom. Likewise, we reach $\widehat{KL}^{\text{eff},a} = 0.01$ with a robustified Gaussian with $c = 4.2$ and a Student $t$ with $\nu = 12.7$ degrees of freedom.

We now illustrate the ability of robust filters to avoid weight-degeneracy problems.

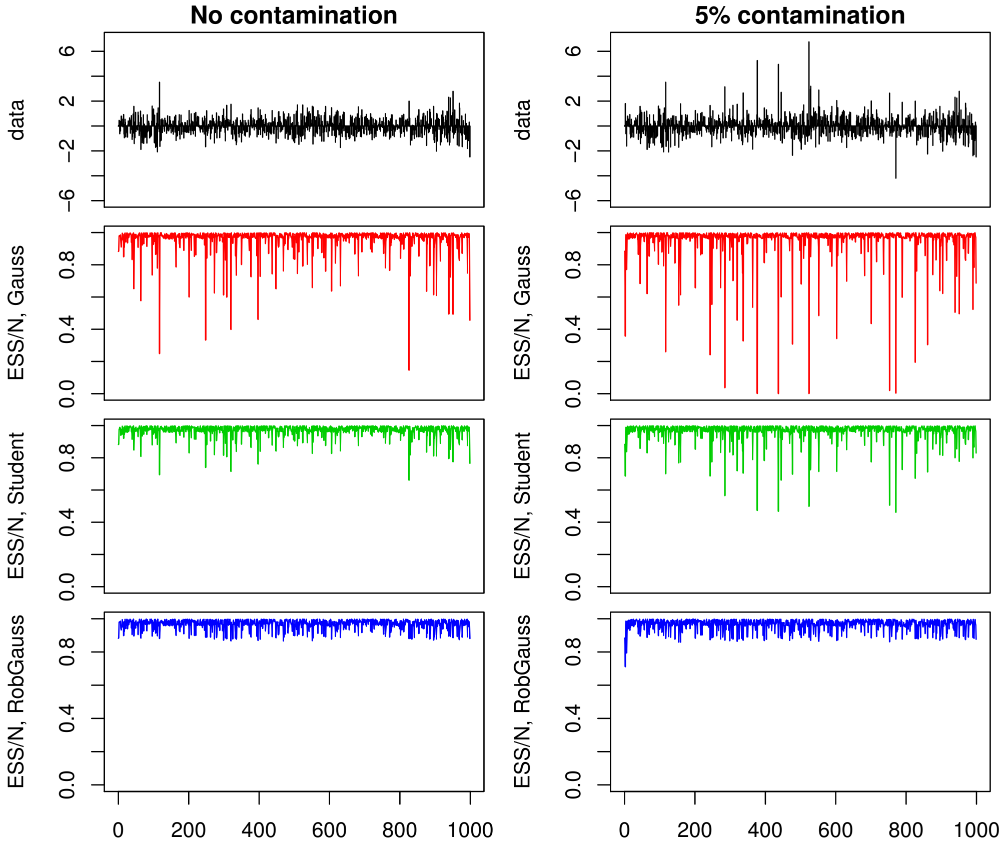

Figure 9: Proportion of effective particle sizes under uncontaminated (left panel) and 5%-contaminated (right panel) SV simulations, using robustified Gaussian and Student $t$ particle filters with $c = 2.8$ and $\nu = 4.9$. Sample size is $T = 10^3$ and filter size is $N = 10^6$.

As in Kong, Liu and Wong (1994), we use the effective sample size

$$ESS_t = \left\{ \sum_{i=1}^{N} \left( \frac{\omega_t^{(i)}}{\sum_{j=1}^{N} \omega_t^{(j)}} \right)^2 \right\}^{-1}, \tag{4.16}$$

where $\{\omega_t^{(i)}\}_{i=1}^N$ are second-step importance weights defined in Section 3.6.2. The top panels of Figure 9 illustrate an uncontaminated (left panel) and a 5% RO-contaminated (right panel) SV simulation with contamination (4.2) and $\eta = 4$ of size $T = 10^3$. The three bottom panels in each column of Figure 9 illustrate the proportion of effective particles $ESS_t/N$ in standard (second panel), Student $t$ (third panel) and robustified Gaussian (fourth panel) particle filters. The robustified Gaussian and Student $t$ filters are based on $N = 10^6$ particles and tuning constants $c = 2.8$ and $\nu = 4.9$. The standard particle filter encounters weight-degeneracy problems even when there is no contamination in the underlying observation process and reaches a minimal proportion of alive particles of 14.59% in the uncontaminated case and 0.04% in the contaminated cases. Robust filters, and in particular the robustified Gaussian, are much less affected by weight degeneracy problems. The proportion of alive particles in the Student $t$ filter is always above 66.09% in the uncontaminated and 46.13% in the contaminated case; while in the robustified Gaussian filter, it is always above 86.61% under no contamination and 71.09% under contamination.

The robust particle filter can be used to estimate the parameters of the SV model. The boxplots in Figure 10 contain 200 estimates of $a$, $b$ and $\sigma$ obtained by maximizing the log-likelihood approximated by standard, Student $t$ and robustified Gaussian particle filters. Data are generated from uncontaminated (left panel) and 5%-contaminated (right panel) SV models with contamination (4.2) and $\eta = 4$. The sample size is $T = 10^3$ and the particle filter size is $N = 10^6$. When there is no contamination in the data, the three methods estimate the model parameters equally well. However, huge accuracy gains are obtained using robust methods in the contaminated case.

Additional information is provided by Figure 11 which shows the smooth behavior of the log-likelihood function obtained through the robust particle filter when the data are 5%-contaminated and model specifications, data and filter size are the same as in Figure 10. This should be contrasted with the rugged behavior of the same

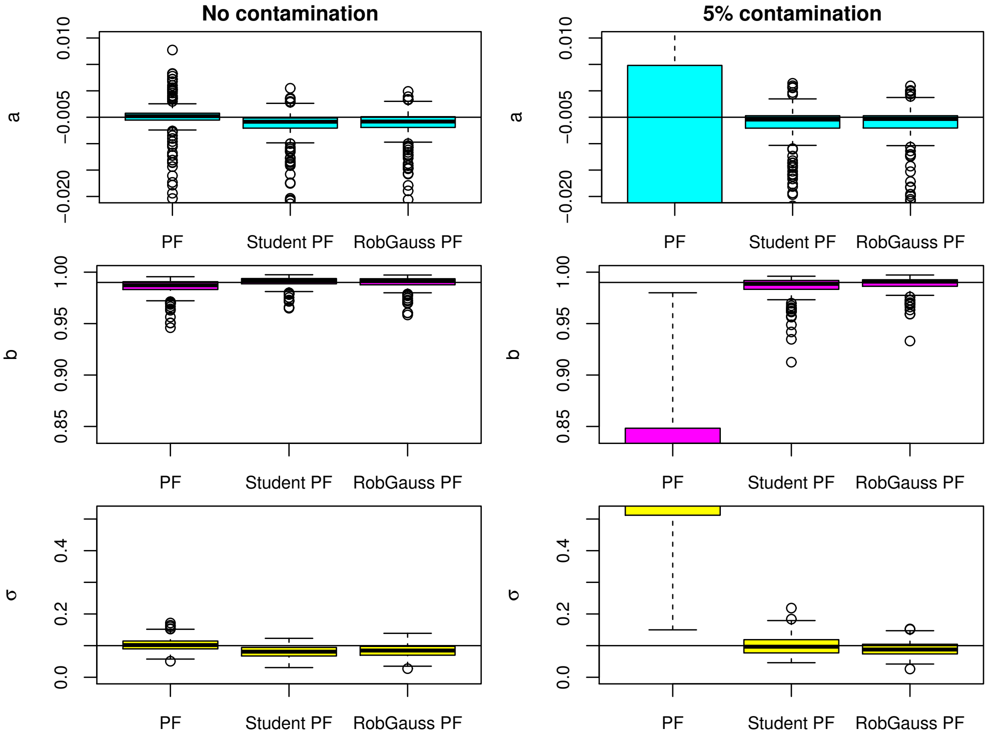

Figure 10: Particle filter ML estimates of the SV parameters using standard, a Student $t$ with $\nu = 4.9$-robustified and a truncated density-robustified filter with $c = 2.8$ without (left panel) and with 5% contamination with $\eta = 4$ (right panel). Sample size is $T = 10^3$ and filter size is $N = 10^6$.

function for $a$ and $b$ obtained using the standard particle filter (which would lead in addition to numerical problems).

Finally, we can compare the robust method with the auxiliary particle filter (Pitt and Shephard, 1999), an algorithm that samples states in Step 1 from a distribution other than the kernel. In work not reported here, we found that auxiliary filters produce very similar results to the ones obtained in Figures 9–11 with the standard bootstrap filter. The explanation is that the auxiliary filter's importance weights are proportional to the model's observation densities and are therefore highly sensitive to outliers, just like the weights of the standard filter. In future work, it might be

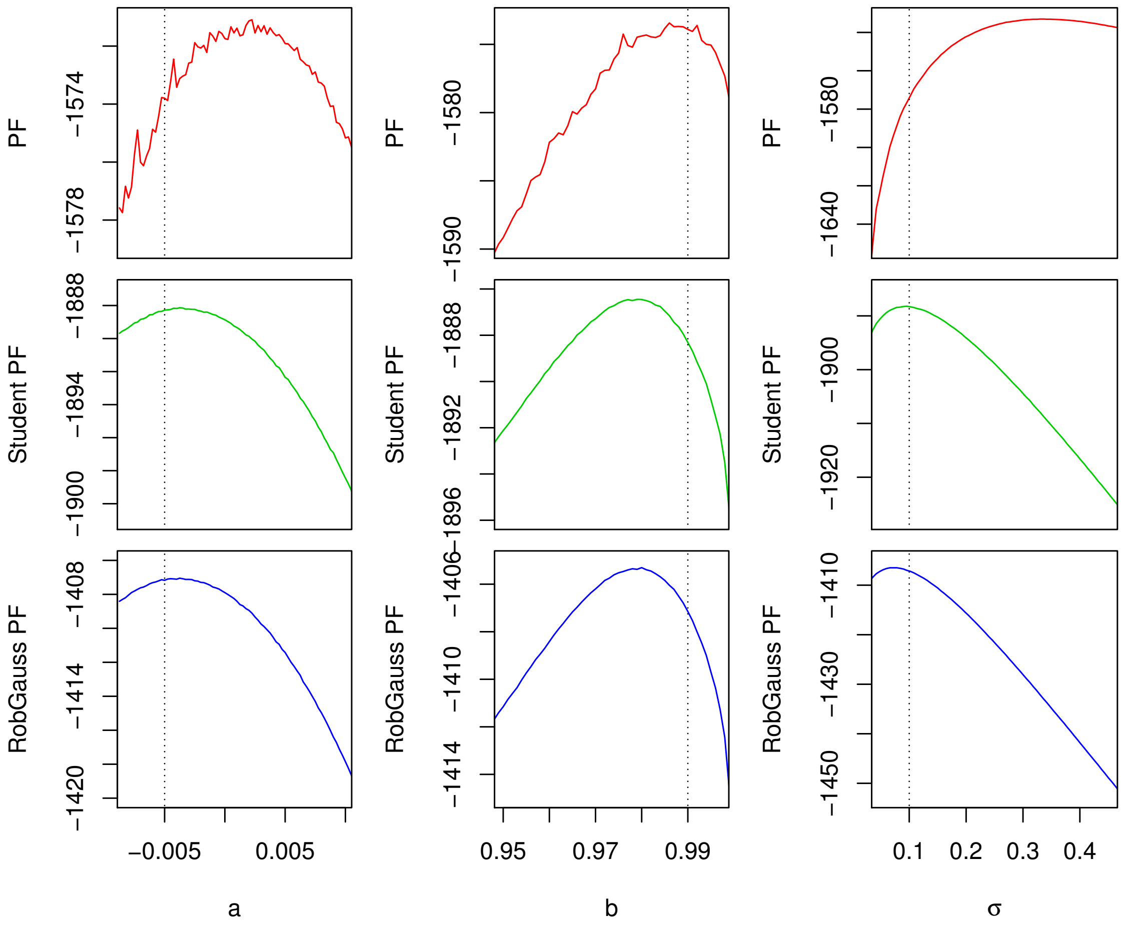

Figure 11: Objective functions cuts of the particle filter estimated likelihood function for the estimation of the SV parameters using standard and robust filters with $c = 2.8$ and $\nu = 4.9$ with 5% contamination with $\eta = 4$. Sample size is $T = 10^3$ and filter size is $N = 10^6$. Real parameter values are shown with dotted lines.

interesting to robustify auxiliary filters along the lines of this paper, which might bring about further improvements in efficiency.

# 5 Conclusion

In this paper, we have developed a general class of robust filters based on the methodology of robust statistics. Our method applies to any observation process $y_t$ that is defined on Euclidean space, is driven by a latent Markov state $x_t$, and has a smooth observation density $f(y_t|x_t, Y_{t-1})$. The robust filter provides accurate state and parameter inference, even if the model is slightly misspecified. This novel approach is based on the impact function, which we have defined as the sensitivity of the state distribution $\lambda(x_t|Y_t)$ with respect to new data. By Bayes' rule, the impact function also measures the sensitivity of the observation density $f(y_t|x_t, Y_{t-1})$. Traditional Bayesian filters have unbounded impact functions, even in the simplest examples. This leads to define a filter to be robust if it has a bounded impact function.

We have illustrated the good performance of the new method in a number of examples. First, we have shown its excellent precision in filtering applications. The robust filter entails only modest efficiency costs in the absence of contamination, but achieves large efficiency gains in the presence of modest contamination. In particular, the robust filter is as accurate as the robustified Kalman filter used in the literature, even though our filter is fully general and does not rely on the Gaussian linear structure of the state space. Second, because it is less sensitive to outliers, the robust method naturally solves the degeneracy problem that plagues the bootstrap particle filter and its many extensions. Third, the robust particle filter provides a highly accurate likelihood-based model selection method, with or without contamination. It is also useful for parameter estimation, for instance in the context of single- or multi-factor stochastic volatility models. Fourth, the robust filter generates smoother probability estimates and smoother likelihood functions than its standard counterpart. In complex settings in which a particle filter must be used, the robust filter permits to use less particles than its traditional counterpart. Computational speed can therefore be an additional benefit of our approach.

The paper offers several directions for future research. One would like to better understand the computational gains that robustness permits to achieve for particle filtering. It would also be useful to extend the approach to a number of closely related filtering environments. For instance, one would like to construct a robust version of particle filters developed for systems in which the observation density is not available in closed form (Calvet and Czellar, 2012). Application to online parameter learning (Storvik, 2002) are also envisioned and will be the subject of future research.

# A General Assumptions

**Assumption 1 (Contamination)** *The class of admissible disturbances is*

$$\mathcal{V}_k(y_t; Y_{t-1}) = \{v_t \in \mathbb{R}^p \text{ } s.t. \text{ } \|v_t\| \leq k\|y_t - \mathbb{E}(y_t|Y_{t-1})\|\}.$$

The constant $k$ defines an upper bound for the ratio between the disturbance (noise) and the possible deviations of the observation at time $t$ (signal) from its conditional expected value given the past.

We now list assumptions on the observation density

**Assumption 2 (Observation density)** *For every instant $t$, state $x_t$ and observation history $Y_{t-1}$, the observation density $f(y_t|x_t, Y_{t-1})$ is strictly positive and twice continuously differentiable at every $y_t \in \mathbb{R}^p$. Furthermore, the observation density has a finite differential entropy: $-\int_{\mathbb{R}^p} f(y_t|x_t, Y_{t-1}) \log[f(y_t|x_t, Y_{t-1})]dy_t \in \mathbb{R}$.*

**Assumption 3 (Critical region)** *For every $c \in \mathbb{R}_+$ and for every $x_t$, $Y_{t-1}$, and $z \in \mathbb{R}^p$, the critical region*

$$\left\{ y \in \mathbb{R}^p \text{ s.t. } \left\| \frac{\partial \log f(y|x_t, Y_{t-1})}{\partial y_t} \right\| \|y - \mathbb{E}(y_t|Y_{t-1})\| = c \right\} ,$$

*intersects the segment $[\mu(x_t), z]$ finitely many times.*

Assumptions 2 and 3 are satisfied by many standard models, such as the Gaussian and Student distributions considered in the main text.

For the optimality results (Proposition 6), we rely on the following condition.

**Assumption 4 (Regular region)** *For every $x_t$, $Y_{t-1}$ and $c > 1$, the region*

$$\mathcal{R} = \left\{ y \in \mathbb{R} \text{ s.t. } \left| \frac{d \log f(y|x_t, Y_{t-1})}{dy_t} \right| |y - \mathbb{E}(y_t|Y_{t-1})| \leq c \right\} , \tag{A.1}$$

*is an interval.*

# B Proofs

## Proof of Proposition 1

Bayes' rule implies that $\lambda(x_t|Y_{t-1})/\lambda(x_t|y_t, Y_{t-1}) = f(y_t|Y_{t-1})/f(y_t|x_t, Y_{t-1})$, and

$$\begin{aligned} KL\big[\lambda(x_t|Y_{t-1}), \lambda(x_t|y_t, Y_{t-1})\big] =& \mathbb{E}_{\lambda(x_t|Y_{t-1})} \left[ \log \frac{f(y_t|Y_{t-1})}{f(y_t|x_t, Y_{t-1})} \right] \\ =& \log f(y_t|Y_{t-1}) - \mathbb{E}_{\lambda(x_t|Y_{t-1})}[\log f(y_t|x_t, Y_{t-1})]. \end{aligned}$$

Hence, the impact function is

$$I(y_t; \lambda, Y_{t-1}, v_t) = v_t' \left\{ \frac{\partial \log f(y_t|Y_{t-1})}{\partial y_t} - \mathbb{E}_{\lambda(x_t|Y_{t-1})} \left[ \frac{\partial \log f(y_t|x_t, Y_{t-1})}{\partial y_t} \right] \right\} ,$$

where

$$\begin{aligned}\frac{\partial \log f(y_t|Y_{t-1})}{\partial y_t} &= \frac{1}{f(y_t|Y_{t-1})}\frac{\partial}{\partial y_t}\left\{\mathbb{E}_{\lambda(x_t|Y_{t-1})}[f(y_t|x_t,Y_{t-1})]\right\} \\ &= \frac{1}{f(y_t|Y_{t-1})}\mathbb{E}_{\lambda(x_t|Y_{t-1})}\left[f(y_t|x_t,Y_{t-1})\frac{\partial \log f(y_t|x_t,Y_{t-1})}{\partial y_t}\right] \\ &= \mathbb{E}_{\lambda(x_t|Y_t)}\left[\frac{\partial \log f(y_t|x_t,Y_{t-1})}{\partial y_t}\right] .\end{aligned}$$

## Proof of Proposition 2

By Proposition 1, the impact function $|I(y_t;\lambda,Y_{t-1},v_t)|$ is bounded above by

$$\|v_t\|\left[\mathbb{E}_{\lambda(x_t|Y_t)}\left(\left\|\frac{\partial \log f(y_t|x_t,Y_{t-1})}{\partial y_t}\right\|\right)+\mathbb{E}_{\lambda(x_t|Y_{t-1})}\left(\left\|\frac{\partial \log f(y_t|x_t,Y_{t-1})}{\partial y_t}\right\|\right)\right].$$

Since $\|v_t\| \leq k\|y_t - \mathbb{E}(y_t|Y_{t-1})\|$, we conclude that $|I(y_t;\lambda,Y_{t-1},v_t)| \leq 2kc$.

## Proof of Proposition 3

Since the construction is based on truncation, we begin with the following definition.

**Definition 3 (Semi-differentiability)** *Consider the function $U : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D}$ is a subset of $\mathbb{R}^p$. Let $y^*$ denote an interior element of $\mathcal{D}$. We say that $U$ is semi-differentiable at $y^*$ if for every $v \in \mathbb{R}^p$,*

$$\partial_v U(y^*) = \lim_{\varepsilon\to 0^+}\frac{U(y^*+\varepsilon v)-U(y^*)}{\varepsilon}$$

*exists as a real number.*

We can easily show the following lemma.

**Lemma 1** *Consider a continuous function*

$$\begin{aligned} u: \quad [0,1] \times \mathcal{D} &\longrightarrow \mathbb{R} \\ (s,y) &\longmapsto u(s,y), \end{aligned}$$

*where $\mathcal{D} \subset \mathbb{R}^p$. Assume that $y^*$ is in the interior of $\mathcal{D}$ and that for every $s \in [0,1]$, the function $u(s,\cdot)$ is semi-differentiable at $y^*$. Furthermore, there exists a nonnegative measurable function $m(s)$, $\int_0^1 m(s)ds < \infty$, such that for all $\varepsilon > 0$ and $v \in \mathbb{R}^p$,*

$$|u(s, y^* + \varepsilon v) - u(s, y^*)| \leq \varepsilon \|v\| m(s). \tag{B.1}$$

*Then the function*

$$U(y) = \int_0^1 u(s,y)ds$$

*is semi-differentiable at $y^*$, and $\partial_v U(y^*) = \int_0^1 \partial_v u(s,y^*)ds$ for every $v \in \mathbb{R}^p$.*

**Proof of Lemma 1.** We know that $\varepsilon^{-1}[u(s,y^* + \varepsilon v) - u(s,y^*)]$ converges pointwise to $\partial_v u(s,y^*)$ as $\varepsilon \to 0$. Lebesgue's dominated convergence theorem implies that

$$\frac{U(y^* + \varepsilon v) - U(y^*)}{\varepsilon} = \int_0^1 \frac{u(s,y^* + \varepsilon v) - u(s,y^*)}{\varepsilon} ds \underset{\varepsilon \to 0}{\longrightarrow} \int_0^1 \partial_v u(s,y^*)ds$$

□

Consider $M : \mathbb{R}^p \longrightarrow [0,\infty]$ defined by $M(y) = c/\|y - \mu_t\|$. The function

$$g(y) = \frac{\partial \log f}{\partial y}(y|x_t, Y_{t-1}) \min\left(1; \frac{M(y)}{\|(\partial \log f/\partial y)(y|x_t, Y_{t-1})\|}\right)$$

is finite and semi-differentiable at every $y \in \mathbb{R}^p$.

**Lemma 2** *The equation*

$$\frac{\partial \log \tilde{f}}{\partial y}(y|x_t, Y_{t-1}) = g(y) \tag{B.2}$$

*has a unique solution* $\tilde{f}$ *such that* $\tilde{f}[\mu(x_t)|x_t, Y_{t-1}] = f[\mu(x_t)|x_t, Y_{t-1}]$. *Furthermore,* $\tilde{f}$ *belongs to* $C^1(\mathbb{R}^p)$ *and satisfies (3.2).*

**Proof of Lemma 2.** Let $\tilde{f}$ denote a solution to (B.2). The function $\xi(s) = \log \tilde{f}[\mu(x_t) + s(y - \mu(x_t))|x_t, Y_{t-1}]$ has derivative $\xi'(s) = [y - \mu(x_t)]' g[\mu(x_t) + s(y - \mu(x_t))]$ . Since

$$\log \tilde{f}(y|x_t, Y_{t-1}) = \xi(1) = \xi(0) + \int_0^1 \xi'(s)ds$$

we infer that equation (B.2) has at most one solution up to a normalizing constant.

We now check that the proposed solution (3.2) is indeed correct. The auxiliary function $u(s, y) = [y - \mu(x_t)]' g[\mu(x_t) + s(y - \mu(x_t))]$ is semi-differentiable with respect to $y$ and satisfies

$$\partial_v u(s, y) = v' g[\mu(x_t) + s(y - \mu(x_t))] + s\, [y - \mu(x_t)]' \partial_v g[\mu(x_t) + s(y - \mu(x_t))]\,,$$

for every $s \in [0, 1]$ and $y, v \in \mathbb{R}^p$. It follows from Assumption 2 that condition (B.1) holds. Hence for all $v \in \mathbb{R}^p$,

$$\partial_v \log \tilde{f}(y|x_t, Y_{t-1}) = \int_0^1 \partial_v u(s, y)ds.$$

The function $h(s) = s\, v' g[\mu(x_t) + s(y - \mu(x_t))]$ is continuous, semi-differentiable and

$$\partial_+ h(s) = v'\, g[\mu(x_t) + s(y - \mu(x_t))] + s\, v'\, \partial_{y-\mu(x_t)} g[\mu(x_t) + s(y - \mu(x_t))].$$

By Assumption 3, there are at most finitely many points $s_1, \dots, s_J$ where $s \mapsto g[\mu(x_t) + s(y - \mu(x_t))]$ is not differentiable. For any $s \notin \{s_1, \dots, s_J\}$,

$$\partial_{y-\mu(x_t)} g[\mu(x_t) + s(y - \mu(x_t))]'v = \partial_v g[\mu(x_t) + s(y - \mu(x_t))]'[y - \mu(x_t)]$$

and therefore $\partial_+ h(s) = \partial_v u(s, y)$. We infer that

$$\partial_v \log \tilde{f}(y) = \int_0^1 \partial_v u(s, y) ds = h(1) - h(0) = g(y)'v.$$

Since $\partial_v \log \tilde{f}(y)$ is linear in $v$, the function $\log \tilde{f}(y)$ is differentiable and solves equation (B.2). □

Lemma 2 implies that the proposition holds.

## Constants in the Robustified Univariate Gaussian of Section 3.2

- If $c > [\mu(x_t) - \mu_t]^2/[2\sigma(x_t)]^2$, the constants in equation (3.8) are $D_{1,t}(x_t) = |y^*_- - \mu_t|^c f_{\mathrm{N}}[y^*_-; \mu(x_t), \sigma(x_t)^2]$ and $D_{2,t}(x_t) = |y^*_+ - \mu_t|^c f_{\mathrm{N}}[y^*_+; \mu(x_t), \sigma(x_t)^2]$.
- If $c \leq [\mu(x_t) - \mu_t]^2/[2\sigma(x_t)]^2$ and $\mu(x_t) \leq \mu_t$, the constants in equation (3.9) are $C_{1,t}(x_t) = |y^*_- - \mu_t|^c f_{\mathrm{N}}[y^*_-; \mu(x_t), \sigma(x_t)^2]$, $C_{2,t}(x_t) = |z^*_- - \mu_t|^{-c} f_{\mathrm{N}}[z^*_-; \mu(x_t), \sigma(x_t)^2]$, $C_{3,t}(x_t) = C_{2,t}(x_t)|z^*_+ - \mu_t|^c / f_{\mathrm{N}}[z^*_+; \mu(x_t), \sigma(x_t)^2]$ and $C_{4,t}(x_t) = C_{3,t}(x_t)|y^*_+ - \mu_t|^c f_{\mathrm{N}}[y^*_+; \mu(x_t), \sigma(x_t)^2]$.

## Proof of Proposition 4

We observe that $g(y) = h_{c/\|y-\mu_t\|}\left\{-\Sigma(x_t)^{-1}[y - \mu(x_t)]\right\}$, or equivalently

$$g(y) = -\Sigma(x_t)^{-1}[y - \mu(x_t)] \min\left\{1; \frac{c}{\|y - \mu_t\| \, \|\Sigma(x_t)^{-1}[y - \mu(x_t)]\|}\right\}.$$

Our main task is therefore to compute

$$\begin{aligned} & \int_0^1 g\left\{\mu(x_t)+s[y_t-\mu(x_t)]\right\}'[y_t-\mu(x_t)]ds \\ = & -[y_t-\mu(x_t)]'\Sigma(x_t)^{-1}[y_t-\mu(x_t)] \\ & \times\int_0^1 s\min\left[1;\frac{c}{s\left\|\mu(x_t)-\mu_t+s[y_t-\mu(x_t)]\right\|\left\|\Sigma(x_t)^{-1}\left[y_t-\mu(x_t)\right]\right\|}\right]ds. \end{aligned}$$

To simplify notation, we consider $\tilde{\mu}_t=\mu_t-\mu(x_t)$, $\tilde{y}_t=y_t-\mu(x_t)$, and

$$\begin{aligned} \Omega(\tilde{y}_t;0,1) &= \int_0^1 g\left[\mu(x_t)+s\tilde{y}_t\right]'\tilde{y}_t ds \\ &= -\tilde{y}_t'\Sigma(x_t)^{-1}\tilde{y}_t\int_0^1 s\min\left[1;\frac{c}{s\left\|s\tilde{y}_t-\tilde{\mu}_t\right\|\left\|\Sigma(x_t)^{-1}\tilde{y}_t\right\|}\right]ds. \end{aligned}$$

*Threshold points.* The no-truncation condition is equivalent to

$$s^2\left\|s\tilde{y}_t-\tilde{\mu}_t\right\|^2\left\|\Sigma(x_t)^{-1}\tilde{y}_t\right\|^2\leq c^2.$$

The quartic equation

$$\chi(s)=s^2\left\|\Sigma(x_t)^{-1}\tilde{y}_t\right\|^2\left(s^2\left\|\tilde{y}_t\right\|^2-2s\tilde{\mu}_t'\tilde{y}_t+\left\|\tilde{\mu}_t\right\|^2\right)-c^2=0$$

has at most four real roots. One of the roots is negative since $\chi(0)=-c^2<0$ and $\lim_{s\to-\infty}\chi(s)=+\infty$. So the equation $\chi(s)=0$ has at most three roots in $[0,1]$.

*Integration.* We now compute

$$\Omega(\tilde{y}_t;a,b)=-\tilde{y}_t'\Sigma(x_t)^{-1}\tilde{y}_t\int_a^b s\min\left[1;\frac{c}{s\left\|s\tilde{y}_t-\tilde{\mu}_t\right\|\left\|\Sigma(x_t)^{-1}\tilde{y}_t\right\|}\right]ds.$$

Consider an interval $[a, b]$ over which there is no truncation. Then

$$\Omega(\tilde{y}_t; a, b) = -\tilde{y}_t' \Sigma(x_t)^{-1} \tilde{y}_t (b^2 - a^2)/2.$$

If instead there is truncation on $[a, b]$, then

$$\Omega(\tilde{y}_t; a, b) = -c \frac{\tilde{y}_t' \Sigma(x_t)^{-1} \tilde{y}_t}{\|\Sigma(x_t)^{-1} \tilde{y}_t\|} \int_a^b \frac{ds}{\|s\tilde{y}_t - \tilde{\mu}_t\|}.$$

Note that $\|s\tilde{y}_t - \tilde{\mu}_t\| = \sqrt{s^2 \|\tilde{y}_t\|^2 - 2(\tilde{\mu}_t' \tilde{y}_t)s + \|\tilde{\mu}_t\|^2}$. Hence

$$\Omega(\tilde{y}_t; a, b) = -c \frac{\tilde{y}_t' \Sigma(x_t)^{-1} \tilde{y}_t}{\|\Sigma(x_t)^{-1} \tilde{y}_t\|} \int_a^b \left[ \left( s \|\tilde{y}_t\| - \frac{\tilde{\mu}_t' \tilde{y}_t}{\|\tilde{y}_t\|} \right)^2 + \frac{\|\tilde{y}_t\|^2 \|\tilde{\mu}_t\|^2 - (\tilde{\mu}_t' \tilde{y}_t)^2}{\|\tilde{y}_t\|^2} \right]^{-1/2} ds.$$

We infer that the function $\Omega$ reduces to

$$\Omega(\tilde{y}_t; a, b) = -c\beta(y_t, x_t) \left[ \mathrm{sgn} \left( s \|\tilde{y}_t\| - \frac{\tilde{\mu}_t' \tilde{y}_t}{\|\tilde{y}_t\|} \right) \log \left| s \|\tilde{y}_t\| - \frac{\tilde{\mu}_t' \tilde{y}_t}{\|\tilde{y}_t\|} \right| \right]_{s=a}^{s=b} \tag{B.3}$$

if $|\tilde{\mu}_t' \tilde{y}_t| = \|\tilde{y}_t\| \|\tilde{\mu}_t\|$, and is otherwise given by:

$$\Omega(\tilde{y}_t; a, b) = -c\beta(y_t, x_t) \left[ \log \left( s\|\tilde{y}_t\| - \frac{\tilde{\mu}_t' \tilde{y}_t}{\|\tilde{y}_t\|} + \|s\tilde{y}_t - \tilde{\mu}_t\| \right) \right]_{s=a}^{s=b}. \tag{B.4}$$

We easily verify that for $y_t(s) = \mu(x_t) + s[y_t - \mu(x_t)]$:

$$s \|\tilde{y}_t\| - \frac{\tilde{\mu}_t' \tilde{y}_t}{\|\tilde{y}_t\|} = \frac{\tilde{y}_t'(s\tilde{y}_t - \tilde{\mu}_t)}{\|\tilde{y}_t\|} = \frac{[y_t(s) - \mu(x_t)]' \, [y_t(s) - \mu_t]}{\|y_t(s) - \mu(x_t)\|}. \tag{B.5}$$

We plug (B.5) into (B.3) and (B.4) and conclude that the proposition holds.

## Critical Roots of the Robustified Gaussian

If $y_t \neq \mu(x_t)$, solving (3.11) is equivalent to finding the roots of

$$\tilde{\chi}(s) = s^4 - 2\frac{\tilde{\mu}_t'\tilde{y}_t}{\|\tilde{y}_t\|^2}s^3 + \frac{\|\tilde{\mu}_t\|^2}{\|\tilde{y}_t\|^2}s^2 - \frac{c^2}{\|\Sigma(x_t)^{-1}\tilde{y}_t\|^2\|\tilde{y}_t\|^2} \,.$$

The derivative of $\tilde{\chi}(s)$ is $\tilde{\chi}'(s) = 2s\left[2s^2 - 3\tilde{\mu}_t'\tilde{y}_t s/\|\tilde{y}_t\|^2 + \|\tilde{\mu}_t\|^2/\|\tilde{y}_t\|^2\right]$. Consider the discriminant $\Delta = \|\tilde{y}_t\|^{-4}\left[9(\tilde{\mu}_t'\tilde{y}_t)^2 - 8\|\tilde{\mu}_t\|^2\|\tilde{y}_t\|^2\right]$. We note that $\tilde{s}_0 = 0$ is a root of $\tilde{\chi}'$. If $\Delta \geq 0$, the real numbers $\tilde{s}_\pm = [3(\tilde{\mu}_t'\tilde{y}_t)/\|\tilde{y}_t\|^2 \pm \sqrt{\Delta}]/4$ are also zeros of $\tilde{\chi}'$, and have the property that $\mathrm{sgn}(\tilde{s}_-) = \mathrm{sgn}(\tilde{s}_+) = \mathrm{sgn}(\tilde{\mu}_t'\tilde{y}_t)$. This suggests the following algorithm for computing the roots of $\tilde{\chi}$.

1. If $\tilde{\mu}_t'\tilde{y}_t < 0$ or $\tilde{s}_- > 1$ or $\Delta \leq 0$, then $\tilde{\chi}$ has at most one root in the open interval $(0, 1)$, which is obtained by dichotomy if $\tilde{\chi}(1) > 0$.

2. Otherwise, the computation proceeds as follows. If $\tilde{\chi}(\tilde{s}_-) > 0$, there is a root in $[0, \tilde{s}_-]$ which is obtained by dichotomy. For the investigation of the roots in $[\tilde{s}_-, 1]$, we consider the following subcases.

   (a) If $\tilde{s}_+ > 1$: there is a root in $[\tilde{s}_-, 1]$ if $\mathrm{sgn}[\tilde{\chi}(\tilde{s}_-)] \neq \mathrm{sgn}[\tilde{\chi}(1)]$ and it is obtained by dichotomy.

   (b) If $\tilde{s}_+ \leq 1$, we implement the following two steps (which are based on conditions that are not mutually exclusive).

      - If $\mathrm{sgn}[\tilde{\chi}(\tilde{s}_-)] \neq \mathrm{sgn}[\tilde{\chi}(\tilde{s}_+)]$ there is a root in $[\tilde{s}_-, \tilde{s}_+]$, which we compute by a dichotomy.
      - If $\mathrm{sgn}[\tilde{\chi}(\tilde{s}_+)] \neq \mathrm{sgn}[\tilde{\chi}(1)]$, there is a root in $[\tilde{s}_+, 1]$, which we compute by a dichotomy.

This procedure completes Step 1 of the robustified Gaussian algorithm.

## Proof of Proposition 5

By Remark 2, the nonnormalized robustified density coincides with the model's observation density if $\|y_t - \mu_t\| \leq \sqrt{c}\sigma(x_t)$, and is given by (3.12) otherwise. We also recall that the surface of the unit sphere in $\mathbb{R}^p$ is $2\pi^{p/2}/\Gamma(p/2)$. Hence, if $c > p$,

$$\begin{aligned}\int_{\mathbb{R}^p} \tilde{f}(y|x_t, Y_{t-1})dy &= \int_0^{\sqrt{c}} \frac{e^{-r^2/2}}{(2\pi)^{p/2}} \frac{2\pi^{p/2} r^{p-1}}{\Gamma(p/2)} dr + \int_{\sqrt{c}}^{+\infty} \frac{e^{-c/2}}{(2\pi)^{p/2}} \frac{c^{c/2}}{r^c} \frac{2\pi^{p/2} r^{p-1}}{\Gamma(p/2)} dr \\ &= 1 - \frac{1}{2^{p/2-1}\Gamma(p/2)} \int_{\sqrt{c}}^{\infty} r^{p-1} e^{-r^2/2} dr + \frac{e^{-c/2} c^{c/2}}{2^{p/2-1}\Gamma(p/2)} \int_{\sqrt{c}}^{\infty} \frac{1}{r^{c-p+1}} dr \\ &= 1 - \frac{1}{2^{p/2-1}\Gamma(p/2)} \int_{\sqrt{c}}^{\infty} r^{p-1} e^{-\rho^2/2} dr + \frac{e^{-c/2} c^{p/2}}{2^{p/2-1}(c-p)\Gamma(p/2)}.\end{aligned}$$

Let $I_p(x) = \int_x^\infty r^{p-1} e^{-r^2/2} dr$. We integrate $I_p(x)$ by parts and obtain:

$$I_p(x) = x^{p-2} e^{-x^2/2} + (p-2) I_{p-2}(x).$$

We note that $I_1(x) = \sqrt{2\pi}[1 - \Phi(x)]$, and $I_2(x) = e^{-x^2/2}$. Hence

$$I_{2m}(x) = e^{-x^2/2}\,(m-1)!\,2^{m-1} \sum_{i=0}^{m-1} \frac{1}{i!} \left(\frac{x^2}{2}\right)^i,$$

$$I_{2m+1}(x) = e^{-x^2/2} \frac{(2m-1)!}{2^{m-2}(m-1)!} \sum_{i=0}^{m-1} \frac{2^i (i+1)!}{(2i+2)!} x^{2i+1} + \frac{(2m-1)!}{2^{m-1}(m-1)!} \sqrt{2\pi}[1 - \Phi(x)].$$

We know that

$$\int_{\mathbb{R}^p} \tilde{f}(y|x_t, Y_{t-1})dy = 1 + \frac{1}{2^{p/2-1}\Gamma(p/2)} \left[\frac{e^{-c/2} c^{p/2}}{c-p} - I_p(\sqrt{c})\right].$$

We plug in $I_p(\sqrt{c})$ and conclude that the first part of the proposition holds.

We next show:

**Lemma 3** *If $\tilde{f}(y_t|x_t, Y_{t-1}) \geq f(y_t|x_t, Y_{t-1})$ for all $x_t$, $y_t$, the Kullback-Leibler divergence in (3.5) satisfies: $KL_t^{\textit{eff}} \leq -\mathbb{E}_{f(y_t|Y_{t-1})}\left(\log\{\mathbb{E}[B_t(x_t)|Y_t]\}\right)$.*

**Proof of Lemma 3.** Equations (3.3) and (3.4) imply that

$$KL\left[f(y_t|Y_{t-1}), \hat{f}(y_t|Y_{t-1})\right] = \mathbb{E}_{f(y_t|Y_{t-1})} \log\left\{\frac{f(y_t|Y_{t-1})}{\mathbb{E}_{\lambda(x_t|Y_{t-1})}[B_t(x_t)\tilde{f}(y_t|x_t, Y_{t-1})]}\right\}.$$

Since $\tilde{f}(y_t|x_t, Y_{t-1}) \geq f(y_t|x_t, Y_{t-1})$ for all $x_t$, $y_t$, the denominator satisfies

$$\mathbb{E}_{\lambda(x_t|Y_{t-1})}[B_t(x_t)\tilde{f}(y_t|x_t, Y_{t-1})] \geq \mathbb{E}_{\lambda(x_t|Y_{t-1})}[B_t(x_t)\underbrace{f(y_t|x_t, Y_{t-1})}_{\frac{\lambda(x_t|Y_t)f(y_t|Y_{t-1})}{\lambda(x_t|Y_{t-1})}}] = f(y_t|Y_{t-1})\mathbb{E}[B_t(x_t)|Y_t]\,,$$

and we conclude that the lemma holds. □

If $\|y_t - \mu_t\| \geq \sqrt{c}\sigma(x_t)$, the ratio

$$\frac{f(y_t|x_t, Y_{t-1})}{\tilde{f}(y_t|x_t, Y_{t-1})} = \underbrace{\frac{e^{c/2}}{(c\sigma(x_t)^2)^{c/2}}}_{\equiv A_t} \frac{\|y_t - \mu_t\|^c}{e^{\|y_t-\mu_t\|^2/(2\sigma(x_t)^2)}} = A_t \frac{z_t^c}{e^{z_t^2/(2\sigma(x_t)^2)}}$$

is a decreasing function of the distance $z_t = \|y_t - \mu_t\|$, since

$$\frac{\partial[f(y_t|x_t, Y_{t-1})/\tilde{f}(y_t|x_t, Y_{t-1})]}{\partial z_t} = A_t z_t^{c-1} \frac{c - z_t^2/\sigma(x_t)^2}{e^{z_t^2/(2\sigma(x_t)^2)}} \leq 0\,.$$

The robustified observation density therefore satisfies assumption in Lemma 3 and the second part of the proposition holds.

## Proof of Proposition 6

Throughout this proof we shorten notation to $h(y) = h(y|x_t, Y_{t-1})$. The set of admissible functions, which we denote by $\mathcal{H}$, consists of every $h \in \mathcal{PC}^1(\mathbb{R})$ such that $\int_{\mathbb{R}} h(y)dy = 1$, $h(y) > 0$,

$$-(y - \mu_t)\partial_+ h(y) - c\,h(y) \leq 0, \tag{B.6}$$

$$-(y - \mu_t)\partial_+ h(y) + c\,h(y) \leq 0, \tag{B.7}$$

$$-(y - \mu_t)\partial_- h(y) - c\,h(y) \leq 0, \tag{B.8}$$

$$-(y - \mu_t)\partial_- h(y) + c\,h(y) \leq 0, \tag{B.9}$$

for all $y \in \mathbb{R}$.

**Lemma 4** *The set $\mathcal{H}$ is non-empty and convex.*

**Proof of Lemma 4** Under Assumption 4 and the condition $c > 1$, the nonnormalized robustified density $\tilde{f}$ has a finite integral over the real line. The normalized density $\hat{f}(y|x_t, Y_{t-1})$ is well-defined and belongs to $\mathcal{H}$, so $\mathcal{H}$ is not empty.

Given $h_1$ and $h_2$ in $\mathcal{H}$ and $s_1, s_2 \in [0, 1]$, $s_1 + s_2 = 1$, the convex combination $h = s_1 h_1 + s_2 h_2$ satisfies the linear constraints and we conclude that $h \in \mathcal{H}$. □

The optimization problem can be rewritten as:

$$\min_{h \in \mathcal{H}} W(h)$$

where $W(h) = \int_{\mathbb{R}} f(y|x_t, Y_{t-1}) \log[f(y|x_t, Y_{t-1})/h(y)]dy$. For every $h \in \mathcal{H}$, the functional $W(h) = KL[f(y|x_t, Y_{t-1}), h(y)]$ is well-defined and belongs to $[0, +\infty]$. The functional is strictly convex, and standard arguments (Ekeland and Témam, 1987) imply that the following results hold.

**Lemma 5.** *The program $\min_{h \in \mathcal{H}} W(h)$ has at most one solution. A function $h \in \mathcal{H}$*

*is a local optimum of $W$ if and only if satisfies the first-order condition:*

$$W'(h)(k-h) \geq 0 \tag{B.10}$$

*for every $k \in \mathcal{H}$. Furthermore, any local optimum of $W$ on $\mathcal{H}$ is a global optimum.*

We now use the calculus of variations to provide a tractable version of (B.10).

**Euler-Lagrange Equation.** To simplify notation, we observe that the constraint can be rewritten as:

$$\operatorname{sgn}(\mu_t - y)\frac{d}{dy}[|y-\mu_t|^c h(y)] \leq 0 \tag{B.11}$$

for all $y \in \mathbb{R}$. Let

$$G(y,h,h') = -f(y|x_t,Y_{t-1})\log[h(y)] + \lambda h(y) + \nu(y)\operatorname{sgn}(\mu_t - y)\frac{d}{dy}[|y-\mu_t|^c h(y)].$$

We note that

$$\begin{aligned}\frac{\partial G}{\partial h} &= -\frac{f(y|x_t,Y_{t-1})}{h(y)} + \lambda - c\,|y-\mu_t|^{c-1}\,\nu(y), && \text{(B.12)}\\ \frac{\partial G}{\partial h'} &= \nu(y)\operatorname{sgn}(\mu_t - y)|y-\mu_t|^c. && \text{(B.13)}\end{aligned}$$

The Euler-Lagrange equation

$$\frac{\partial G}{\partial h} = \frac{d}{dy}\left[\frac{\partial G}{\partial h'}\right]$$

is equivalent to

$$-\frac{f(y|x_t,Y_{t-1})}{h(y)} + \lambda + \operatorname{sgn}(y-\mu_t)\,\nu'(y)\,|y-\mu_t|^c = 0 \tag{B.14}$$

for every $y \in \mathbb{R}$.

For given threshpoints $\underline{y}^*$ and $\bar{y}^*$, we consider the function $h^*(y) = f^*(y|x_t, Y_{t-1})$ and the normalizing constant $B_t(x_t)$ defined by (3.20) and (3.21). We let

$$\nu^*(y) = \begin{cases} \int_y^{+\infty} \frac{\lambda^* h^*(z) - f(z)}{h^*(z)\ |z-\mu_t|^c} dz & \text{if} \quad y \geq \bar{y}^*, \\ 0 & \text{if} \quad y \in [\underline{y}^*, \bar{y}^*], \\ \int_{-\infty}^{y} \frac{\lambda^* h^*(z) - f(z)}{h^*(z)\ |z-\mu_t|^c} dz & \text{if} \quad y < \underline{y}^*. \end{cases} \tag{B.15}$$

The functions $h^*$ and $\nu^*$ and the Lagrange multiplier $\lambda^* = 1/B_t(x_t)$ satisfy the Euler-Lagrange equation. Since

$$\nu^*(\underline{y}^*) = \frac{\lambda^*}{(\mu_t - \underline{y}^*)^{c-1}} \left[ \frac{1}{c-1} - \frac{F(\underline{y}^*|x_t, Y_{t-1})}{(\mu_t - \underline{y}^*)\, f(\underline{y}^*|x_t, Y_{t-1})} \right]$$

the condition $\nu^*(\underline{y}^*) = 0$ holds if and only if $\underline{y}^*$ satisfies (3.22). Similarly,

$$\nu^*(\bar{y}^*) = \frac{\lambda^*}{(\bar{y}^* - \mu_t)^{c-1}} \left[ \frac{1}{c-1} - \frac{1 - F(\bar{y}^*|x_t, Y_{t-1})}{(\bar{y}^* - \mu_t)\, f(\bar{y}^*|x_t, Y_{t-1})} \right]$$

is equal to zero if and only if (3.23) holds.

**Computing the Gâteaux derivative.** The Gâteaux derivative of the functional $W$ at $h^*$ is given by

$$W'(h^*)(h - h^*) = \int_{\mathbb{R}} -\frac{f(y|x_t, Y_{t-1})}{h^*(y)} [h(y) - h^*(y)] dy$$

for every $h \in \mathcal{H}$. We note that

$$W'(h^*)(h-h^*) \;\geq\; \int_{\mathbb{R}} \left[-\frac{f(y|x_t, Y_{t-1})}{h^*(y)} + \lambda^*\right] [h(y) - h^*(y)]dy \\ + \int_{\mathbb{R}} \nu(y)\text{sgn}(\mu_t - y)\frac{d}{dy}[|y-\mu_t|^c[h(y) - h^*(y)]dy.$$

We infer from (B.15) that

$$0 \leq \nu(y)|y-\mu_t|^c\,[h(y) - h^*(y)] \leq \frac{\lambda^*}{c-1}|y-\mu_t|\,[h(y)-h^*(y)],$$

and therefore $\lim_{|y|\to+\infty} \nu(y)|y-\mu_t|^c\, h^*(y) = 0$. We can therefore integrate by parts on the real line:

$$\int_{\mathbb{R}} \nu(y)\text{sgn}(\mu_t - y)\frac{d}{dy}\{|y-\mu_t|^c[h(y)-h^*(y)]\}dy = \int_{\mathbb{R}} \nu'(y)\text{sgn}(y-\mu_t)|y-\mu_t|^c[h(y)-h^*(y)]dy.$$

We plug this relation into (B.16) and infer from (B.14) that $W'(h^*)(h-h^*) \geq 0$. We conclude from Lemma 5 that the first part of the Proposition holds.

**General program**. We now consider the program under the general condition $\int_{\mathbb{R}} h(y)dy = m$. The homogeneity of the constraints implies that $m\,h^*$ belongs to the admissible set of the general program. Furthermore since $W(mh) = W(h) - \log(m)$ for every admissible function, the density $h$ solves (3.24) if and only if $m\,h$ solves the general program. We conclude that the second part of the Proposition holds.

# References

Calvet, L. E., and V. Czellar (2012). Tracking Beliefs: Accurate Methods for Approximate Bayesian Computation Filtering. Working paper, HEC Paris.

Calvet, L. E., and A. J. Fisher (2001). Forecasting Multifractal Volatility. *Journal of Econometrics* **105**, 27–58.

Calvet, L. E., and A. J. Fisher (2008). *Multifractal Volatility: Theory, Forecasting and Pricing.* Elsevier – Academic Press.

Calvet, L. E., Fisher, A. J., and S. Thompson (2006). Volatility Comovement: A Multifrequency Approach. *Journal of Econometrics* **131**, 179–215.

Cappé, O., Moulines, E., and T. Rydén (2005). *Inference in Hidden Markov Models.* Springer-Verlag, New York.

Chib, S., Nardari, F., and N. Shephard (2006). Analysis of High-Dimensional Multivariate Stochastic Volatility Models. *Journal of Econometrics* **134**, 341–371.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* Springer-Verlag, New York.

Doucet, A., De Freitas, N. and Gordon, N.J. (editors) (2001). *SMC Methods in Practice.* Springer-Verlag, New York.

Doucet, A., and A. M. Johansen (2011). A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later. *Oxford Handbook of Nonlinear Filtering*, Oxford University Press.

Ekeland, I., and R. Témam (1987). *Convex Analysis and Variational Problems.* Society for Industrial and Applied Mathematics.

Gordon, N., Salmond, D., and A. F. Smith (1993). Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. *IEE Proceedings F* **140**, 107–113.

Hamilton, J. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* **57**, 357–384.

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

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

Huber, P. J. (1981). *Robust Statistics.* Wiley, New York.

Huber, P. J., and E. M. Ronchetti (2009). *Robust Statistics.* 2nd edition, Wiley, New York.

Johannes, M., and N. Polson (2009), "Particle Filtering," in *Handbook of Financial Time Series*, ed. by T. Andersen, R. A. Davis, J.-P. Kreiss, and Th. Mikosch, 1015–1028. Springer.

Kalman, R. E. (1960), A New Approach to Linear Filtering and Prediction Problems, *Journal of Basic Engineering, Transactions of the ASME Series D* **82**, 35–45.

Kong, A., Liu, J. S., and W. H. Wong (1994), Sequential Imputations and Bayesian Missing Data Problems, *Journal of the American Statistical Association* **89**, 278–288.

Lindgren, G. (1978). Markov Regime Models for Mixed Distributions and Switching Regressions. *Scandinavian Journal of Statistics* **5**, 81–91.

Lockwood, E. H. (1961). *A Book of Curves.* Cambridge University Press.

Maronna, R. A., Martin, R. D., and V. J. Yohai (2006). *Robust Statistics: Theory and Methods.* Wiley, Chichester.

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

Pitt, M., and N. Shephard (1999). Filtering Via Simulation: Auxiliary Particle Filters. *Journal of the American Statistical Association* **94**, 590–599.

Ruckdeschel, P. (2010a). Optimally (Distributional-) Robust Kalman Filtering. Technical Report, Fraunhofer ITWM Kaiserslautern (Germany), arXiv:1004.3393v1 [math.ST].

Ruckdeschel, P. (2010b). Optimally Robust Kalman Filtering at Work: AO-, IO, and Simultaneously IO- and AO- Robust Filters. Technical Report, Fraunhofer ITWM Kaiserslautern (Germany), arXiv:1004.3895v1 [stat.CO].

Ruckdeschel, P., Spangl, B., and D. Pupashenko (2012). Robust Kalman Tracking and Smoothing with Propagating and Non-propagating Outliers. Technical Report, Fraunhofer ITWM Kaiserslautern (Germany), arXiv:1204.3358v2 [math.ST].

Schick, I. C., and S. K. Mitter (1994). Robust Recursive Estimation in the Presence of Heavy-Tailed Observation Noise. *Annals of Statistics* **22**, 1045–1080.

Storvik, G. (2002). Particle Filters in State Space Models with the Presence of Unknown Static Parameters. *IEEE Transactions on Signal Processing* **50**, 281–289.