State Space Methods for Robust Estimation in AR-Models With External Regressors

Christian Sangiorgio

Diss. ETH No. 16063 Zürich, 2005

Diss. ETH No. 16063

State Space Methods for Robust Estimation in AR-Models With External Regressors

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZURICH

for the degree of
Doctor of Mathematics

presented by
CHRISTIAN SANGIORGIO
Dipl. Math. ETH
born June 20, 1973
citizen of Balerna TI

accepted on the recommendation of
Prof. Dr. H. R. Künsch, examiner
Prof. Dr. F. Hampel, co-examiner
Prof. Dr. W. Stahel, co-examiner

2005

To my parents

La sapienza è figliola della sperienzia.
Leonardo da Vinci

Acknowledgements

During the thesis, I encountered many difficulties. Theory and reality do not always meet themselves.
Many people have contributed with stimulating discussions and direct aid to this thesis. First of all, I would like to thank my supervisor Hans-Ruedi Künsch and the R-guru Martin Mächler for their constant support. Furthermore, my gratitude goes to Werner Stahel, Bruno Tona, Johannes Stähelin and Christina Colberg for providing the Gubrist data and helping me in their analysis. I am in dept with Michael Psarros for introducing me in the C language. A thanks goes to the colleagues at the "Seminar für Statistik" for the warm atmosphere in these years.
Last but not least, I would like to thank all friends and my parents. Only with their constant support and encouragement I have completed this thesis.

Contents

Summary **iii**

Riassunto **v**

1 Introduction **1**

2 Filtering recursion **9**

2.1 Exact filtering recursion 12

2.2 Particle filtering method 15

2.2.1 Rejection method 17

2.2.2 Sampling importance resampling method (SIR) . 20

2.3 Particle filtering recursion with $\widetilde{y}_t$ available 21

2.3.1 Construction of proposal densities $\rho\,(i, z)$ 24

2.3.2 Construction of the distribution $(\tau(i))$ 30

2.3.3 Acceptance probability of the proposed pair (i, z) 31

2.3.4 Construction of majorants 33

2.3.5 Summary of the particle filtering recursion with $\widetilde{y}_t$ available 42

2.4 Particle filtering recursion with missing $\widetilde{y}_t$ 44

3 Smoothing recursion **47**

3.1 Exact density $p\,(z_{1:T}|\widetilde{y}_{1:T})$ 48

3.2 Particle smoothing method 55

3.3 Particle smoothing recursion: method 1 57

3.3.1 Recursion with $\widetilde{y}_t$ available 57

3.3.2 Recursion with missing $\widetilde{y}_t$ 76

3.4 Particle smoothing recursion: method 2 77
3.4.1 Recursion with $\widetilde{y}_t$ available 79
3.4.2 Recursion with missing $\widetilde{y}_t$ 98

4 Parameter estimation 103
4.1 Approximation of the log-likelihood function based on the smoothing sample 105
4.2 MCEM method to approximate the log-likelihood function 108
4.3 Algorithm to compute starting estimates 109
4.3.1 Estimation of hyperparameters 110
4.3.2 Estimation of nuisance parameters 110

5 Examples 119
5.1 State estimates given the parameters 119
5.1.1 Gaussian error distributions without outliers . . 120
5.1.2 Gaussian error distributions with one outlier . . 128
5.2 Parameter estimation 133
5.2.1 Estimates using an AR(1) state process 136
5.2.2 Estimates using an AR(4) state process 143
5.3 Estimation of road traffic emissions 148

6 Robustness of filter and smoother distributions 159

7 Outlook 169

Bibliography 171

Curriculum Vitae 177

Summary

Regression models with time series as both response and explanatory variables play an important role in several fields of science. The easiest time series regression model is built with independent $\mathcal{N}(0, \sigma^2)$ distributed errors. But these error assumptions are not fulfilled in many real situations. In reality, correlation in the random errors is present and/or some outliers are observable.
Many methods have been developed to take into consideration these features. The random errors have been modeled using an AR or ARIMA process. Robust methods have been developed for time series autoregressions, generalizing robust methods applied in the usual regression context. But the proposed procedures have only been partially adapted to the time series situation and a simultaneous treatment of regression models with outliers and correlated errors has been missing.

A promising idea to handle both correlated errors and outliers is to write the time series regression models in state space form. State space models constitute a large and flexible class of models. They consist of an equation which describes the dynamics of the observation sequence (Y_t) using an unobserved first order Markov process (X_t) (the *state process*). For example, the popular Gaussian ARMA model is equivalent to a linear Gaussian state space model.
Parameter estimation and inference for state space models rely on the assessment of the distributions of the unknown states X_t given the observed values $Y_1, \ldots, Y_s$. Three cases are distinguished: the prediction $(t > s)$, the filtering $(t = s)$ and the smoothing $(t < s)$ *state distributions.* General recursive methods for calculating the respective state space distributions have been derived, but the formulae contain convolution integrals. Thus, an easy and closed form for the results is possible only under strict assumptions (Gaussian distributions for both the er-

rors and the starting X_0). But, the resulting methods are not robust and therefore outliers can have disastrous effects. Unfortunately, it is not easy to robustify these methods. On the other hand, it is possible to choose other error distributions than the Gaussian one to model the outliers. Then, approximations are needed in order to compute the state distributions. A new approach which works well also with high dimensional states X_t consists of approximating the state distributions by samples generated using Monte Carlo methods.

In the present thesis, we consider a time series regression model and assume linearity in the error term. Both auto-regressive correlated errors and observation outliers (*additive outliers*) are considered and no constraints are set on the relationship between response and explanatory time series. Moreover, missing values in the response series and/or in the explanatory series are allowed.
The resulting state distributions are computed approximately using Monte Carlo techniques. Special care is used to develop recursive algorithms which are both fast and reliable. In fact, these characteristics are a prerequisite for the estimation of the unknown parameters in the considered time series regression model.
The estimation issue is solved by applying the maximum likelihood method. The resulting estimates are robust thanks to the assumed error distributions. In addition, the maximum likelihood method permits to compute approximate confidence intervals by the usual likelihood procedures. The difficulty with this approach is that the likelihood function cannot be computed in closed form and thus it has to be approximated using again Monte Carlo methods. The developed methods are illustrated with some examples.
Finally, we consider the robustness of filter and smoother distributions to analyse how reliable the developed methods are.

Riassunto

In parecchi campi della scienza si incontrano modelli di regressione con variabili indipendenti e variabile dipendente date da serie temporali. Il modello più semplice in questi casi assume errori indipendenti e distribuiti normalmente con valore atteso zero e varianza σ^2. Purtroppo queste assunzioni per l'errore sono disattese in molte applicazioni reali. Infatti gli errori sono correlati e/o presentano valori anomali ("outliers").
Parecchi metodi sono stati sviluppati per tenere in considerazione queste caratteristiche. Per esempio gli errori sono stati modellati usando processi AR o ARIMA. Inoltre metodi robusti sono stati sviluppati per serie temporali con errori autocorrelati generalizzando i metodi robusti usati nella regressione. Ma le procedure proposte sono state solo parzialmente adattate alle caratteristiche delle serie temporali. L'analisi simultanea di modelli di regressione con errori correlati e valori anomali è ancora agli inizi.

Un' idea promettente per trattare sia gli errori correlati che i valori anomali è di scrivere le regressioni tra serie temporali usando modelli di stato ("state space models"). Questi ultimi costituiscono una classe ampia e flessibile di modelli. Consistono in un' equazione che descrive la dinamica delle osservazioni (Y_t) usando un processo markoviano (X_t) di ordine uno che non è osservabile direttamente (il cosiddetto *processo di stato*). Per esempio il modello ARMA con errori gaussiani, attualmente molto in voga, è equivalente a un modello lineare di stato con errori gaussiani.
Il problema maggiore nell'utilizzare i modelli di stato è dato dal calcolare le distribuzioni degli stati X_t conoscendo le osservazioni $Y_1, \ldots, Y_s$. Tre casi vengono distinti: la distribuzione di stato per la previsione ($t > s$), per il filtro ($t = s$) e per la ricostruzione a posteriori ("smoothing",

$t < s$). Per calcolare queste distribuzioni sono state sviluppate ricorsioni generali che però contengono convoluzioni. Così ricorsioni facili e in forma chiusa possono essere derivate solo nel caso in cui gli errori e la distribuzione iniziale X_0 siano date dalla distribuzione di Gauss. Queste ricorsioni però non sono robuste e così valori anomali nelle osservazioni possono avere un effetto disastroso. Purtroppo non è facile rendere robuste le ricorsioni trovate. D'altra parte, distribuzioni di errori diverse da quella di Gauss possono essere usate per modellare i valori anomali ma con lo svantaggio di dover usare delle approssimazioni per poter calcolare le distribuzioni di stato. Un nuovo approccio che funziona pure per stati X_t con dimensione elevata consiste nell'approssimare le distribuzioni di stato con un campione generato usando metodi di Monte Carlo.

Nella presente tesi consideriamo un modello di regressione tra serie temporali con linearità nei termini d'errore. Sia errori autocorrelati che valori anomali additivi sono presi in considerazione. Inoltre nessuna restrizione viene posta alla funzione che lega la serie temporale dipendente alle serie temporali indipendenti. Alcuni valori delle serie temporali dipendente e/o indipendenti possono mancare.

Le distribuzioni di stato che ne derivano sono calcolate approssimativamente usando metodi di Monte Carlo. Particolare cura è usata per sviluppare algoritmi ricorsivi che siano allo stesso tempo veloci e affidabili. Infatti queste due caratteristiche sono un prerequisito fondamentale per svillupare i metodi di stima dei parametri sconosciuti nel modello considerato.

La stima in sè viene trovata applicando il metodo della massima verosimiglianza ("maximum likelihood method"). Le stime risultanti sono robuste grazie alle distribuzioni dell'errore scelte. Inoltre, il metodo della massima verosimiglianza permette di trovare intervalli di confidenza approssimativi usando le tecniche sviluppate per questo metodo.

La difficoltà in questo approccio sta nel fatto che la funzione di probabilità ("likelihood function") non può essere espressa in forma chiusa per il modello considerato e così deve essere approssimata usando nuovamente metodi di Monte Carlo. I metodi sviluppati vengono poi illustrati tramite alcuni esempi.

Da ultimo, consideriamo la robustezza delle distribuzioni di filtro ed "smoothing" trovate per analizzare quanto affidabili siano i metodi sviluppati.

Chapter 1

Introduction

Regression type models with time series as both response and explanatory variables are common in many fields of science, for example in finance, biology and in many engineering domains. In the easiest time series regression model, it is assumed that random errors are independent and $\mathcal{N}(0, \sigma^2)$ distributed. Frequently, these assumptions are far from being fulfilled. The random errors are correlated and/or some outliers are observable.

The classical way to deal with the dependence among random fluctuations assumes an AR or ARIMA model for the error term. A pioneer work in this field is the paper written by Cochrane and Orcutt (1949). They developed a stepwise method to cope with correlated errors. On the other hand, deviations from the normal distribution assumption for the error term call for robust estimation methods and the respective inference techniques. Many methods exist for i.i.d. heavy-tailed errors and they are well studied, see for example Hampel et al. (1986). Robust methods have also been developed for time series auto-regressions, generalizing robust methods applied in the normal regression framework. For example, Fox (1972) analysed the effects on time series of two outlier types: the outliers in the observations (additive outliers) and the outliers in the innovations (innovation outliers). The former influence the time series only at the times where they arise, and the aim is to attenuate their effects. The latter produce structural changes in the time series, and thus affect the time series also in successive times. In this case, the goal is to follow these changes as fast as possible.

Unfortunately, the introduced methods have only been partially adapted to the time series situation. In addition, a simultaneous treatment of outliers and correlated errors has been missing.

In the early sixties, Kalman (1960) and Kalman and Bucy (1961) introduced a very general model which includes a whole class of special cases: the state space model. This model was primarily introduced for researches in the aerospace domain.
State space models consist of an equation which describes the observation dynamics using an unobserved first order Markov process. Kalman and Bucy considered a linear state space model. Thus, the unobserved Markov process is described by the *state equation*

$$X_t = G_t X_{t-1} + V_t \tag{1.1}$$

and the connection between states and observations is given by the *observation equation*

$$Y_t = H_t X_t + W_t. \tag{1.2}$$

Here, the states (X_t) are k-dimensional, the observations (Y_t) are l-dimensional, (G_t) are $k \times k$-matrices and (H_t) are $l \times k$-matrices. They assumed that (V_t) and (W_t) are two independent normally distributed sequences with means zero and covariance matrices Σ_t and Ω_t, respectively. In addition, they supposed that X_0 followed a normal distribution with mean $m_{0|0}$ and covariance matrix $R_{0|0}$. Kalman and Bucy focused their interest on finding the distribution of X_t given $Y_1, \ldots, Y_s$. They distinguished three cases: the prediction $(t > s)$, the filtering $(t = s)$ and the smoothing $(t < s)$ distribution. Under the above assumptions, prediction, filtering and smoothing distributions are again Gaussian and it suffices to compute the conditional means $m_{t|s}$ and covariances $R_{t|s}$. They derived the well-known filtering and smoothing recursions for conditional means and covariances. Explicitly, the filtering recursion is given by

$$\begin{aligned}
m_{t|t-1} &= G_t m_{t-1|t-1}, && (1.3)\\
R_{t|t-1} &= \Sigma_t + G_t R_{t-1|t-1} G_t^{'}, && (1.4)\\
R_{t|t} &= \left(H_t^{'} \Omega_t^{-1} H_t + R_{t|t-1}^{-1}\right)^{-1} \\
&= R_{t|t-1} - R_{t|t-1} H_t^{'} M_t^{-1} H_t R_{t|t-1},
\end{aligned}$$

$$\begin{aligned} m_{t|t} &= m_{t|t-1} + R_{t|t} H_t^{'} \Omega_t^{-1} \left(y_t - H_t m_{t|t-1} \right) && (1.5) \\ &= m_{t|t-1} + R_{t|t-1} H_t^{'} M_t^{-1} \left(y_t - H_t m_{t|t-1} \right) && (1.6) \\ &= m_{t|t-1} + K_t \left(y_t - H_t m_{t|t-1} \right) && (1.7) \end{aligned}$$

with

$$\begin{aligned} M_t &= \Omega_t + H_t R_{t|t-1} H_t^{'}, \\ K_t &= R_{t|t-1} H_t^{'} M_t^{-1}. \end{aligned}$$

Note that the equations (1.5), (1.6) and (1.7) have the intuitive interpretation "the filter mean is equal to the prediction mean plus a correction term which depends on how much the new observation differs from its prediction". In addition, K_t is the so-called Kalman gain. On the other hand, the smoothing recursion is given by

$$\begin{aligned} m_{t|T} &= m_{t|t} + S_t \left(m_{t+1|T} - m_{t+1|t} \right), \\ R_{t|T} &= R_{t|t} - S_t \left(R_{t+1|t} - R_{t+1|T} \right) S_t^{'} \end{aligned}$$

with

$$S_t = R_{t|t} G_{t+1}^{'} R_{t+1|t}^{-1}.$$

These recursions are very appealing and easy to compute. But they present some disadvantages. First, the observations (Y_t) enter linearly in the computations, see (1.5), (1.6) or (1.7). Thus, the effect of outliers in (Y_t) is not reduced and the consequence can be disastrous. In addition, all covariance matrices are independent of the observations and, in the time invariant case, they converge quickly towards steady values. Then, the confidence intervals have constant widths. Moreover, the closed and easy form of the resulting recursions depends strongly on the assumptions that the errors and the starting X_0 have Gaussian distributions.

State space models have been studied intensely in the last decades. At first, only in engineering domains, later also in statistics. In fact, state space methods appeared in the time series literature only in the seventies (Akaike (1974), Harrison and Stevens (1976)), became established in the eighties and then an intensive research topic in the nineties. One reason of their popularity is given by their high flexibility and thus the wide range of possible applications. For example, the Gaussian

ARMA model is equivalent to a linear Gaussian state space model in the time-invariant case (see for example Wei (1990), Chapter 15, for a simple proof and Akaike (1974) or Hannan and Deistler (1988) for more details). The state space representation of an ARMA model is useful since it permits to handle missing data in a very easy way. In fact, if the observation at time t is missing, it suffices to set $H_t = (0, \ldots, 0)^{'}$. This leads to an easy manner to compute the exact likelihood of an ARMA model in the presence of missing observations, see for example Jones (1980).

Many generalizations of the state space model introduced by Kalman and Bucy have been examined. Some attempts have been made to robustify equation (1.5) (or (1.6) or (1.7)). The easiest idea has been to substitute $y_t - H_t m_{t|t-1}$ by $\psi(y_t - H_t m_{t|t-1})$ for a suitable ψ function. In this way, the influence of observation outliers has been reduced, but the covariance matrices have still been independent of the observations. Masreliez (1975) and Martin and Thomson (1982) proposed approximate estimation methods. Unfortunately, these methods do not permit to estimate the parameters simultaneously. A way out is to use an iterative scheme which alternates between removing the outliers and estimating the parameters from the cleaned data. Kitagawa (1987) developed numerical approximations for the filter recursion. But in general, the numerical integration is difficult in higher dimensions and, additionally, a good choice of the integration knots would presuppose a knowledge of how the filter densities look like. The replacement of Gaussian errors with heavy-tailed errors has also been used to model different types of outliers, see the ideas in Fox (1972). An unusually large value in V_t corresponds to an innovative outlier which also influences successive observations Y_s, $s > t$. On the other hand, an unusually large value in W_t corresponds to an additive outlier which affects only Y_t. The difficulty in this approach is the computation of both the filter and smoothing distributions and the maximum likelihood estimate.

A new approach to state space models has become popular and feasible in the last years thanks to the increased computer performances. The filter and smoother densities are approximated by samples produced with Monte Carlo methods. Then, expectations can be approximated by sample averages, quantiles by corresponding order statistics, etc. Pioneers in this field were Carlin et al. (1992). They proposed to use the standard Gibbs sampler to generate samples from the conditional distribution of $(X_1, \ldots, X_T)$ given $(Y_1, \ldots, Y_T)$. This can be performed sampling from the so-called full conditionals, i.e. the density of X_t given the remaining $(X_{t'})$, $t \neq t'$, and $(Y_1, \dots, Y_T)$. But this approach has two disadvantages. The method is not recursive and thus the samples should be recomputed from the beginning when a new observation becomes available. In addition, the convergence of single update methods is usually slow. A method to improve considerably the convergence speed of the Gibbs sampler for some often used models was proposed by some authors, see for example Frühwirth-Schnatter (1994), Carter and Kohn (1994) or Shephard (1994). The single updates in the Gibbs sampler can be substituted by multiple ones in the same step. This is possible in all models where the state variable can be split in two components, $X_t = (X_{t,1}, X_{t,2})'$, and the sampling of both $(X_{t,1})_{t=0,\dots,T}$ given $((X_{t,2})_{t=0,\dots,T}, (Y_t)_{t=1,\dots,T})$ and $(X_{t,2})_{t=0,\dots,T}$ given $((X_{t,1})_{t=0,\dots,T}, (Y_t)_{t=1,\dots,T})$ is realizable. Thus, the sampling proceeds alternating between simulating from one of the two components while the other is kept fixed. However, the lack of recursivity remains. The idea of a recursive Monte Carlo sampling method goes back to Handschin and Mayne (1969) and Handschin (1970). It was proposed again by Gordon et al. (1993), Isard and Blake (1996) and Kitagawa (1996). It has become now very popular and there is an extensive literature on it, see Doucet (1998) and the book edited by Doucet et al. (2001). This method is often called the *particle filter*.

The goal of the present thesis is to apply recursive Monte Carlo methods to a time series regression model of the form

$$Y_t = f(X_{t,1}, \dots, X_{t,m}; \alpha_1, \dots, \alpha_l) + Z_t + \varepsilon_t, \qquad t = 1, \dots, T \qquad (1.8)$$

with (Y_t) an univariate time series and $f(.)$ a function of the external regressors $(X_{t,1})$, …, $(X_{t,m})$ with hyperparameters α_1, …, α_l. In addition, (Z_t) and (ε_t) are assumed to be two independent sequences where (Z_t) is a Gaussian AR(p) process and (ε_t) are i.i.d. distributed according to a Pearson type VII distribution with mean zero, that is a scaled t-distribution.

Linearity in the error term of (1.8) is required. On the other hand, a simultaneous treatment of correlated errors and additive outliers is considered. The former one is given by the Gaussian AR(p) process (Z_t) and the latter one by the chosen heavy-tailed distribution for (ε_t). No restrictions are set to the function $f(.)$. It may be linear or nonlinear and the number of external regressors may be different from the number of hyperparameters. Moreover, some values may be missing in the observation series and/or the external regressors.

The most interesting problem is to estimate the hyperparameters α and

the nuisance parameters (the parameters characterizing the distributions of (Z_t) and (ε_t) in (1.8)). To this aim, the time series regression model (1.8) is written in state space form. Then, the first topic to discuss will be the inference about the unknown states based on observed values $(Y_t, X_{t,1}, \ldots, X_{t,m})$ and on given parameters in (1.8). In fact, reliable and fast algorithms for filtering and smoothing are a prerequisite for the parameter estimation. Moreover, filtering and smoothing algorithms permit the identification of additive outliers. Thus, the recursions for the filter and smoother densities of the unobserved states are derived. Especially with AR processes of order p greater than one, the smoothing recursion is not straightforward. But, unfortunately, the filter and smoother densities cannot be computed in closed form as a consequence of the chosen non-Gaussian observation error distribution. These densities will be approximated using a sequential Monte Carlo method. In addition, filtering and smoothing recursions are implemented to work also in presence of missing values in the observations (Y_t) and/or in the external regressors $(X_{t,1}), \ldots, (X_{t,m})$. Second, the maximum likelihood method is applied to estimate all (or just a subset) of the unknown parameters in (1.8). The resulting estimates are expected to be robust since the observation error distribution is assumed to be heavy-tailed. Moreover, the maximum likelihood method permits to compute also approximate confidence intervals by the usual likelihood procedures. In general, particular care is required to get fast and reliable maximum likelihood algorithms. The difficulty in this context is that the likelihood function cannot be computed in closed form. Therefore, it has to be approximated with Monte Carlo methods which use samples generated for a given set of parameters, see Hürzeler (1998). Consequently, the approximations of the likelihood are reliable only locally and the maximum likelihood procedures have to be iterated until the estimate convergence. For this reason, it is also necessary to develop an algorithm which computes a good starting estimate of the unknown parameters such that the needed number of iterations is small.
A last point to examine is the robustness of filter and smoother distributions, for example if a set of observations Y_t and/or external regressors goes to infinity. This shows how reliable the developed methods are.

The thesis is structured as follows. In Chapter 2, we will examine the filtering inference about the unknown states based on observed values $(Y_t, X_{t,1}, \ldots, X_{t,m})$ and on given parameters in (1.8). The smoothing inference will be the subject of Chapter 3 and maximum likelihood estimation will be consider in Chapter 4 together with the algorithm to compute starting estimates. Chapter 5 illustrates the developed algorithm with some examples. A comparison with the Kalman algorithm is carried out both on simulation studies and on a real data example. The latter one consists of the estimation of vehicle emission factors and it gave actually the input for this thesis. In Chapter 6, the robustness of filter and smoother distributions is analysed. In Chapter 7, some remarks about the developed algorithms and their computer implementation will conclude the thesis.

Chapter 2

Filtering recursion

The aim of this chapter will be to derive an implementable filtering recursion for the considered time series regression model. In fact, the exact filtering recursion can be found easily, but it is not possible to work directly with it. We will derive an approximate filtering recursion using the Monte Carlo method. In addition, the filtering recursion should be implemented in an efficient way. Two methods are presented to sample from the densities of the approximate filtering recursion. The chosen one will be explained in detail. The illustrative examples are postponed to Chapter 5. In this way, we will compare the results of the filtering recursion with the smoothing ones.

First, we explain the considered model. We examine a (possibly nonlinear) time series regression model where the random errors consist of a linear combination of a Gaussian AR(p) process and a heavy-tailed distributed random variable. With the Gaussian AR(p) process we take into account the possible error correlation and with the heavy-tailed terms the presence of outliers. Explicitly, the considered model is given by

$$Z_t = \varphi_1 Z_{t-1} + \cdots + \varphi_p Z_{t-p} + V_t, \tag{2.1}$$

$$Y_t = f(X_{t,1}, \ldots, X_{t,m}) + Z_t + \varepsilon_t \tag{2.2}$$

with

t : time index: $t \in \{1, \dots, T\}$,

(Y_t) : time series of the observed values,

$(X_{t,1}, \dots, X_{t,m})$: known external (explanatory) regressors, where f is a (possibly nonlinear) function of them,

(Z_t) : stationary Gaussian AR(p) process. The coefficients are $\varphi_1, \dots, \varphi_p$,

(V_t) : state errors. (V_t) are $\overset{\text{i.i.d.}}{\sim}$ $\mathcal{N}(0, \sigma_V^2)$ distributed,

(ε_t) : observation errors. (ε_t) are $\overset{\text{i.i.d.}}{\sim}$ Pearson type VII distributed with parameters m, c and $\xi = 0$.

Remark 2.1 *The general Pearson type VII distribution has a probability density function that can be expressed in the form*

$$p_{VII}\left(m, c, \xi\right)(w) = \frac{\Gamma(m)}{\sqrt{\pi}\ c\,\Gamma(m - 0.5)}\ \frac{1}{\left[1 + \left(\frac{w-\xi}{c}\right)^2\right]^m}. \tag{2.3}$$

It depends on the 3 parameters m, c and ξ where m and c should be strictly positive. Expected value and variance of a Pearson type VII random variable W are given by

$$\mathbf{E}\left[W\right] = \xi, \qquad Var\left(W\right) = \frac{c^2}{2m - 3}. \tag{2.4}$$

The t_ν-distribution is a special case of the Pearson type VII distribution. It is obtained by setting $m = \frac{1}{2}(\nu + 1)$, $c = \sqrt{\nu}$ and $\xi = 0$. Moreover

$$\sqrt{2m - 1}\ \frac{W - \xi}{c} \sim t_{2m-1}.$$

More details on the Pearson type VII distribution can be found in Johnson et al. (1995), Chapter 28.

The key idea is that the equations (2.1) and (2.2) can be interpreted as defining a state space model. In fact, the sequence (Z_t) can play the role of the unobserved state sequence with state evolution given by (2.1). The series (Y_t) is the observation sequence generated according to (2.2). In addition, we should assume that the initial distribution of (Z_t) is known:

$$(Z_{1-p}, \dots, Z_0) \sim p_0(z_{1-p}, \dots, z_0) d\mu(z_{1-p}, \dots, z_0).$$

But, in this chapter and in Chapter 3 we are interested in the inference about the states (Z_t) based on a stretch of both observed values (Y_t) and external regressors $(X_{t,1}, \ldots, X_{t,m})$ for a given model (i.e. all parameters in (2.1) and (2.2) are assumed to be known). Thus, it is convenient to define the new observed sequence $(\widetilde{Y}_t)$ as the model residuals

$$\widetilde{Y}_t := Y_t - f(X_{t,1}, \ldots, X_{t,m}). \tag{2.5}$$

The state space model (2.1) and (2.2) simplifies to the model

$$Z_t = \varphi_1 Z_{t-1} + \cdots + \varphi_p Z_{t-p} + V_t, \tag{2.6}$$
$$\widetilde{Y}_t = Z_t + \varepsilon_t \tag{2.7}$$

with $(Z_{1-p}, \ldots, Z_0)$, (V_t) and (ε_t) distributed as before.

Remark 2.2 *As common in the literature, we refer to the equations (2.1), (2.6) and (2.2), (2.7) as the* state equation *and the* observation equation*, respectively.*

The dependence structures of the state space model (2.6) and (2.7) can be illustrated helpfully with the graph in Figure 2.1. Various conditional independence properties can be easily read from it as we will remark in the next sections.

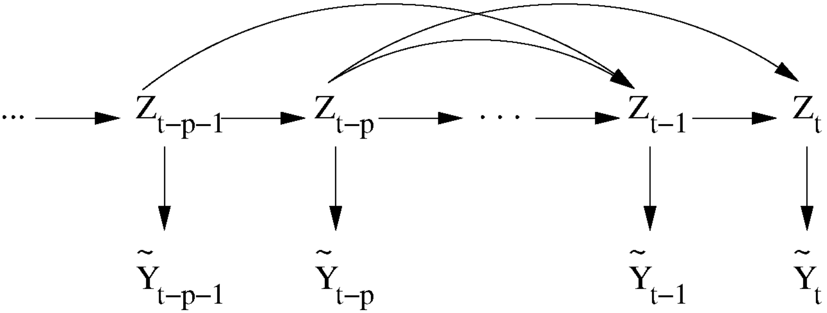

Figure 2.1: *Dependence structures of the state variable Z_t given the observations $\widetilde{Y}_{1:t}$.*

In addition, we set some notations that we will use throughout the whole thesis. We define

- $\widetilde{Y}_{s:t}$, $\widetilde{y}_{s:t}$:
  The set of random variables $\left(\widetilde{Y}_u, s \leq u \leq t\right)$ or realizations $(\widetilde{y}_u, s \leq u \leq t)$.
- $Z_{s:t}$, $z_{s:t}$:
  The set of state random variables $(Z_u, s \leq u \leq t)$ or realizations $(z_u, s \leq u \leq t)$.
- $p(.)$:
  A general probability density function.
- $f_{t|s}(z_t|\widetilde{y}_{1:s}) = p(z_t|\widetilde{y}_{1:s})$:
  The conditional density of z_t given $\widetilde{Y}_{1:s} = \widetilde{y}_{1:s}$. We will consider the cases $s < t$ (prediction) and $s = t$ (filtering) in this chapter. The case $s > t$ (smoothing) will be the topic of Chapter 3.
- $\phi(\mu, \sigma)(z) = \phi_\sigma(z - \mu)$:
  The density of the $\mathcal{N}(\mu, \sigma^2)$ distribution. Since the normal density depends actually on the difference $z - \mu$, it is useful to introduce also the notation $\phi_\sigma(z - \mu)$.
- $p_{VII}(m, c, \xi)(w) = b_{m,c}(w - \xi)$:
  The density of the Pearson type VII distribution with parameters m, c and ξ, see (2.3). Since it depends actually on the difference $w - \xi$, it is useful to introduce also the notation $b_{m,c}(w - \xi)$.

Then, we have for the considered model (2.6) and (2.7):

$$\begin{aligned}
p(\widetilde{y}_t|z_t) &= p_{VII}(m, c, z_t)(\widetilde{y}_t) = p_{VII}(m, c, \widetilde{y}_t)(z_t), \\
p\left(z_t|z_{(t-p):(t-1)}\right) &= \phi\left(\sum_{l=1}^{p} \varphi_l z_{t-l}, \sigma_V\right)(z_t).
\end{aligned}$$

2.1 Exact filtering recursion

As mentioned before, the exact filtering recursion can be derived easily. The easiest case is when (Z_t) is a Gaussian AR(1) process (or in general a first order Markov process). Then, the recursion can be found using the well-known two-step procedure (a propagation step followed by an update step).

Propagation step: From the filter density at time $t-1$, the one step ahead prediction density is obtained by

$$\begin{aligned}
f_{t|t-1}\left(z_t|\widetilde{y}_{1:(t-1)}\right) &:= p\left(z_t|\widetilde{y}_{1:(t-1)}\right) \\
&= \int p\left(z_{t-1}, z_t|\widetilde{y}_{1:(t-1)}\right) dz_{t-1} \\
&= \int p\left(z_t|z_{t-1}, \widetilde{y}_{1:(t-1)}\right) p\left(z_{t-1}|\widetilde{y}_{1:(t-1)}\right) dz_{t-1} \\
&= \int p\left(z_t|z_{t-1}\right) f_{t-1|t-1}\left(z_{t-1}|\widetilde{y}_{1:(t-1)}\right) dz_{t-1} \qquad (2.8) \\
&= \mathbf{E}_{Z_{t-1}|\widetilde{Y}_{1:t-1}}\left[p\left(z_t|z_{t-1}\right)\right]. \qquad (2.9)
\end{aligned}$$

The last equality follows since Z_t is independent of $\widetilde{Y}_{1:(t-1)}$ given Z_{t-1} (see the graph in Figure 2.1).

Update step: From the one step ahead prediction density, the filter density at time t is derived by Bayes' theorem:

$$\begin{aligned}
f_{t|t}\left(z_t|\widetilde{y}_{1:t}\right) &= p\left(z_t|\widetilde{y}_{1:t}\right) \\
&= \frac{p\left(z_t, \widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)}{p\left(\widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)} \\
&= \frac{p\left(\widetilde{y}_t|z_t, \widetilde{y}_{1:(t-1)}\right) p\left(z_t|\widetilde{y}_{1:(t-1)}\right)}{p\left(\widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)} \\
&= \frac{p\left(\widetilde{y}_t|z_t\right) f_{t|t-1}\left(z_t|\widetilde{y}_{1:(t-1)}\right)}{p\left(\widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)}. \qquad (2.10)
\end{aligned}$$

The last equality follows since $\widetilde{Y}_t$ is independent of $\widetilde{Y}_{1:(t-1)}$ given Z_t (see again the graph in Figure 2.1).

Therefore, the starting density of Z_0 and the equations (2.8) and (2.10) permit to find the exact filtering recursion. The filter density at some time t is computed using the filter density at the previous time, starting with the density of Z_0. But it is not straightforward to work with this recursion since the integral in the prediction step (2.8) cannot be computed in closed form for the considered model (2.6) and (2.7).

The question of how to generalize the recursion for AR(p) ($p > 1$) state processes has not been answered yet. Some approaches have been proposed. A first idea considers p-step prediction and filter densities.

Briefly, the p-step ahead prediction density $p\left(z_{(t-p+1):t}|\widetilde{y}_{1:(t-p)}\right)$ is derived. Then, this density is updated using the observations $\widetilde{y}_{(t-p+1):t}$ to find the filter density $p\left(z_{(t-p+1):t}|\widetilde{y}_{1:t}\right)$. The disadvantage is that it is hard to implement such a recursion efficiently since p states $Z_{(t-p+1):t}$ are involved at the same time. A second approach is to consider p-dimensional state vectors $(Z_{(t-p+1):t})$ but with one step prediction and update. A very appealing idea would be to use the first order Markov property of $(Z_{(t-p+1):t})$. In fact, defining $X_t = (Z_t, \ldots, Z_{t-p+1})'$, we have

$$X_t = \begin{pmatrix} \varphi_1 & \varphi_2 & \varphi_3 & \cdots & \varphi_p \\ 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} X_{t-1} + \begin{pmatrix} V_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Unfortunately, the error vector (and thus X_t) does not have a probability density function in $\mathbb{R}^p$ since it has some deterministic components. Then, difficulties arise in implementing the one step filtering recursion for X_t efficiently. The finding of the smoothing distribution presents the same problems since the state densities should be available in analytic form to find the density $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$, see Section 3.1.

We propose a slightly different method. We consider again the p-dimensional state vectors $(Z_{(t-p+1):t})$. But the one step propagation and update are computed directly for $Z_{(t-p+1):t}$, i.e. without using its first order property. In this way, the difficulties mentioned above do not arise. But the resulting filter densities must be approximated by a Monte Carlo method since it is not possible to compute them in closed form. The approximation will be the topic of Section 2.2. Here we derive the recursion explicitly.

Propagation step: From the filter density at time $t-1$, the one step ahead prediction density is obtained by

$$\begin{aligned} & f_{t|t-1}\left(z_{(t-p+1):t}|\widetilde{y}_{1:(t-1)}\right) = p\left(z_{(t-p+1):t}|\widetilde{y}_{1:(t-1)}\right) \\ & = \int p\left(z_{t-p}, z_{(t-p+1):t}|\widetilde{y}_{1:(t-1)}\right) dz_{t-p} \\ & = \int p\left(z_t|z_{(t-p):(t-1)}, \widetilde{y}_{1:(t-1)}\right) p\left(z_{(t-p):(t-1)}|\widetilde{y}_{1:(t-1)}\right) dz_{t-p} \\ & = \int p\left(z_t|z_{(t-p):(t-1)}\right) f_{t-1|t-1}\left(z_{(t-p):(t-1)}|\widetilde{y}_{1:(t-1)}\right) dz_{t-p}. \end{aligned}$$

The last equality follows as in the case with $p = 1$: Z_t is independent of $\widetilde{Y}_{1:(t-1)}$ given $Z_{(t-p):(t-1)}$ (see the graph in Figure 2.1).

Update step: The filter density at time t is derived from the one step ahead prediction density by applying Bayes' theorem:

$$\begin{aligned}
f_{t|t}\left(z_{(t-p+1):t}|\widetilde{y}_{1:t}\right) &= p\left(z_{(t-p+1):t}|\widetilde{y}_{1:t}\right)\\
&= \frac{p\left(z_{(t-p+1):t}, \widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)}{p\left(\widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)}\\
&= \frac{p\left(\widetilde{y}_t|z_{(t-p+1):t}, \widetilde{y}_{1:(t-1)}\right)p\left(z_{(t-p+1):t}|\widetilde{y}_{1:(t-1)}\right)}{p\left(\widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)}\\
&= \frac{p\left(\widetilde{y}_t|z_t\right)f_{t|t-1}\left(z_{(t-p+1):t}|\widetilde{y}_{1:(t-1)}\right)}{p\left(\widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)}.
\end{aligned}$$

The last equality follows since $\widetilde{Y}_t$ is independent of both $\widetilde{Y}_{1:(t-1)}$ and $Z_{(t-p+1):(t-1)}$ given Z_t (see again the graph in Figure 2.1).

Putting the two steps together, we have

$$\begin{aligned}
f_{t|t}\left(z_{(t-p+1):t}|\widetilde{y}_{1:t}\right) = \frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)}\cdot&\\
\cdot\int p\left(z_t|z_{(t-p):(t-1)}\right)f_{t-1|t-1}\left(z_{(t-p):(t-1)}|\widetilde{y}_{1:(t-1)}\right)dz_{t-p}&.
\end{aligned} \tag{2.11}$$

Thus, the filtering recursion follows. Note, however, that the integral cannot be interpreted as expected value with respect to $Z_{(t-p):(t-1)}|\widetilde{Y}_{1:(t-1)}$ in general, compare with the special case $p = 1$ in (2.9).

2.2 Particle filtering method

The integral in (2.11) cannot be computed in analytic form for the considered model (2.6) and (2.7). Therefore, we should approximate it to have an implementable filtering recursion. The key idea is to approximate the filter density $f_{t-1|t-1}\left(z_{(t-p):(t-1)}|\widetilde{y}_{1:(t-1)}\right)$ by its discrete density. I.e.,

$$f_{t-1|t-1}\left(z_{(t-p):(t-1)}|\widetilde{y}_{1:(t-1)}\right) \approx \frac{1}{N}\sum_{i=1}^{N}\Delta\left(z^{(i)}_{(t-p):(t-1)}\right)$$

where $\left(z^{(i)}_{(t-p):(t-1)}\right)$, $i = 1, \ldots, N$, is a sample of the random vector $Z_{(t-p):(t-1)}|\widetilde{Y}_{1:(t-1)}$ and $\Delta(x)$ is the Dirac density in point x. Then, (2.11) can be approximated by

$$\begin{aligned} &f_{t|t}\left(z_{(t-p+1):t}|\widetilde{y}_{1:t}\right) \approx \\ &\approx \frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)}\frac{1}{N}\sum_{i=1}^{N}\int p\left(z_t|z_{(t-p):(t-1)}\right)\Delta\left(z^{(i)}_{(t-p):(t-1)}\right)dz_{t-p} \\ &= \frac{1}{N}\frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|\widetilde{y}_{1:(t-1)}\right)}\sum_{i=1}^{N}p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)\Delta\left(z^{(i)}_{(t-p+1):(t-1)}\right). \qquad (2.12)\end{aligned}$$

Thus, we can derive an approximate filtering recursion if we are able to generate a sample $(z^{(l)}_{(t-p+1):t})$ from (2.12) using the previous filter sample $(z^{(i)}_{(t-p):(t-1)})$. To this aim, we note that the mixture density (2.12) is the marginal distribution of the random variable $(I, Z_{(t-p+1):t})|\widetilde{Y}_{1:t}$ with respect to $Z_{(t-p+1):t}$. Thus, a sample from (2.12) can be found generating first a sample $(z^{(l)}_t)$ from the density $\widehat{f}_{t|t}\left(z_t|\widetilde{y}_{1:t}\right)$ defined by

$$\widehat{f}_{t|t}\left(z_t|\widetilde{y}_{1:t}\right) \;\propto\; p\left(\widetilde{y}_t|z_t\right)\sum_{i=1}^{N}p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right) \qquad (2.13)$$

with $(z^{(i)}_{(t-p):(t-1)})$ as above. (Note that the denominator in (2.12) is just the normalizing constant and thus it is omitted in (2.13).) Then, we look for the indices l of the densities $p\left(z_t|z^{(l)}_{(t-p):(t-1)}\right)$ in (2.13) which are used to generate $(z^{(l)}_t)$ to recover the other components $(z^{(l)}_{(t-p+1):(t-1)})$ and find the sample $(z^{(l)}_{(t-p+1):t})$.

In the rest of the section, we present a brief overview of two general methods to accomplish the sampling from (2.13). The discussion applies to any densities. The sampling method adapted to our aim and optimized for the model (2.6) and (2.7) is presented later, see Sections 2.3 and 2.4. The overview is borrowed from Künsch (2003). Following his notation we introduce:

Definition 2.1 *The density*

$$p\left(z_t\right) = \frac{1}{N}\sum_{i=1}^{N} p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right) \tag{2.14}$$

is called the prior whereas the density $p\left(\widetilde{y}_t|z_t\right)$ *is called the likelihood.*

2.2.1 Rejection method

The rejection method for sampling from the density $\widehat{f}_{t|t}\left(z_t|\widetilde{y}_{1:t}\right)$ produces values according to a proposal $\rho\left(z_t\right)$ and then accepts the generated $Z_t = z_t$ with probability

$$\pi\left(z_t\right) = \frac{p\left(\widetilde{y}_t|z_t\right)\sum_{i=1}^{N} p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right)}{\rho\left(z_t\right)\ M}. \tag{2.15}$$

M is the normalising constant of this expression or an upper bound for it:

$$M \geq \sup_{z_t} \frac{p\left(\widetilde{y}_t|z_t\right)\sum_{i=1}^{N} p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right)}{\rho\left(z_t\right)}.$$

The average acceptance probability of the rejection method is given by

$$\int \rho\left(z_t\right)\pi\left(z_t\right)dz_t = \frac{\sum_{i=1}^{N}\int p\left(\widetilde{y}_t|z_t\right)p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right)dz_t}{M}.$$

The most obvious choice for the proposal $\rho\left(z_t\right)$ is the prior (2.14). Then, the evaluation of the acceptance probabilities $\pi\left(z_t\right)$ is easy as long as $p\left(\widetilde{y}_t|z_t\right)$ is bounded. With the smallest value of M it follows:

$$\pi\left(z_t\right) = \frac{p\left(\widetilde{y}_t|z_t\right)}{\sup_{z_t} p\left(\widetilde{y}_t|z_t\right)}$$

and the average acceptance probability is

$$\int \rho\left(z_t\right)\pi\left(z_t\right)dz_t = \frac{\sum_{i=1}^{N}\int p\left(\widetilde{y}_t|z_t\right)p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right)dz_t}{N\ \sup_{z_t} p\left(\widetilde{y}_t|z_t\right)}.$$

This average acceptance probability is low if the likelihood is more informative (concentrated) than the prior or if the likelihood and the prior are in conflict.

Remark 2.3 *Sampling according to the prior (2.14) is carried out in two steps. First, an index I is chosen uniformly from $\{1,\ldots,N\}$ and then the variable Z_t is generated according to $p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)$ with $I = i$. The densities $p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)$ need not to be available in analytic form. Only the sampling from them should be possible.*

Clearly, other proposal distributions than the prior are possible and they can lead to higher acceptance rates. But the computation of a good upper bound M becomes usually more difficult. Moreover, we face the complication that the determination of the acceptance probabilities involves a sum over i, which should be avoided to increase the speed of the calculations.
Pitt and Shephard (1999) proposed to consider explicitly the index I to solve at least the last problem. I.e., we can generate first the auxiliary index I according to a distribution $(\tau(i))$ and then the variable Z_t according to a density $\rho(i, z_t)$ given $I = i$. We accept the sampled pair (i, z_t) with probability

$$\pi(i, z_t) = \frac{p(\widetilde{y}_t|z_t)\, p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)}{\tau(i)\rho(i, z_t)\;\; M}. \tag{2.16}$$

M is again the normalizing constant of this expression or an upper bound for it:

$$M \geq \sup_{i,z_t} \frac{p(\widetilde{y}_t|z_t)\, p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)}{\tau(i)\rho(i, z_t)}.$$

Since the distribution of the accepted pairs (I, Z_t) is given by

$$\frac{p(\widetilde{y}_t|z_t)\, p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)}{\sum_{i=1}^N \int p(\widetilde{y}_t|z_t)\, p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right) dz_t},$$

the marginal distribution of Z_t is exactly $\widehat{f}_{t|t}(z_t|\widetilde{y}_{1:t})$. Thus, if the pair (i, z_t) is accepted, we simply discard the auxiliary index i and keep z_t. Otherwise, we generate a new pair (i, z_t).
The crucial point in the implementation of this idea is the choice of both the proposal distribution $(\tau(i))$ and the densities $\rho(i, z_t)$. For example, if we take

$$\tau(i) = \frac{1}{N} \qquad \text{and} \qquad \rho(i, z_t) = p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right),$$

we obtain the usual rejection algorithm discussed before. But we can try to increase the acceptance rate by other choices. The following lemma is the cornerstone for the optimal choice of both the distribution $(\tau(i))$ and the densities $\rho(i, z_t)$. First, since the index i runs over a finite set, M can be written as

$$M = \max_i \frac{M_i}{\tau(i)} \qquad \text{with} \qquad M_i \geq \sup_{z_t} \frac{p(\widetilde{y}_t|z_t)\, p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)}{\rho(i, z_t)}.$$

Lemma 2.1 *For a given choice of densities $\rho(i, z_t)$ and bounds M_i, the average acceptance probability is maximal for $\tau(i) \propto M_i$.*

Proof: The Proof is taken from Künsch (2003).
The average acceptance probability is

$$\sum_i \int \tau(i)\rho(i, z_t)\, \pi(i, z_t)\, dz_t = \frac{1}{M} \sum_i \int p(\widetilde{y}_t|z_t)\, p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right) dz_t$$
$$= \left(\max_i \frac{M_i}{\tau(i)}\right)^{-1} \sum_i \int p(\widetilde{y}_t|z_t)\, p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right) dz_t.$$

Since the term $\sum_i \int p(\widetilde{y}_t|z_t)\, p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right) dz_t$ is independent of $\tau(i)$, the average acceptance probability is maximal if and only if $\max_i \frac{M_i}{\tau(i)}$ is minimal. But

$$\max_i \frac{M_i}{\tau(i)} = \left(\sum_j \tau(j)\right) \cdot \max_i \frac{M_i}{\tau(i)} = \sum_j \left(\tau(j) \max_i \frac{M_i}{\tau(i)}\right) \geq \sum_j M_j.$$

The term $\sum_j M_j$ does not depend on $\tau(i)$, anymore. It follows that the minimal value of $\max_i \frac{M_i}{\tau(i)}$ is reached if and only if we have the equality, i.e. if and only if $\frac{M_i}{\tau(i)}$ is constant. □

Remark 2.4 *If $\rho(i, z_t) = p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)$ in Lemma 2.1, the optimal $\tau(i)$'s are constant. This is somewhat surprising. In fact, one could conjecture that it would be better to give higher probability to those indices i for which the mass of $p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)$ is close to* $\arg\sup_{z_t} p(\widetilde{y}_t|z_t)$.

Lemma 2.1 indicates not only how the optimal distribution $(\tau(i))$ can be found once the densities $\rho(i, z_t)$ are known. It also gives a method to construct the optimal densities $\rho(i, z_t)$. In fact, we see from its proof that all M_i's should be small to have a high acceptance probability. Thus, each $\rho(i, z_t)$ should be a good proposal distribution for the density

$$\frac{p(\widetilde{y}_t|z_t)\, p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right)}{\int p(\widetilde{y}_t|z_t)\, p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right) dz_t}. \tag{2.17}$$

Ideally, we would choose this density as $\rho(i, z_t)$. But then, M_i must be close to the normalizing constant $\int p(\widetilde{y}_t|z_t)\, p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right) dz_t$ which is typically not available in closed form. In practice, we approximate the density (2.17) and we choose the approximating density as proposal $\rho(i, z_t)$. In this way, we obtain nearly optimal proposal densities $\rho(i, z_t)$.

2.2.2 Sampling importance resampling method (SIR)

This method generates particles $(z_t^{(k)}; 1 \leq k \leq R)$ according to a chosen proposal $\rho(z_t)$. Then it selects a sample of the desired size N from these particles with inclusion probabilities

$$\pi\left(z_t^{(k)}\right) \propto \frac{p\left(\widetilde{y}_t|z_t^{(k)}\right) \sum_{i=1}^{N} p\left(z_t^{(k)}|z_{(t-p):(t-1)}^{(i)}\right)}{\rho\left(z_t^{(k)}\right)}.$$

Resampling need not to be made at random. There are other methods with reduced variability, for example stratification.

The standard choice for the proposal $\rho(z_t)$ is again the prior (2.14). This was originally proposed by Gordon et al. (1993). Situations with a low acceptance rate in the rejection sampling method typically also have heavily unequal sampling probabilities $\pi\left(z_t^{(k)}\right)$. Thus, many ties are present in the resulting sample. If R is chosen much bigger than N, the number of ties will be smaller, but this at the expense of longer computations. Note that the rejection method is an automatic way of choosing R such that all ties are avoided. In cases where all $p\left(.|z_{(t-p):(t-1)}^{(i)}\right)$ have their main mass in a region where the likelihood is flat and small, the sampling importance resampling method can be much faster than the rejection method and still give approximately equal weights to all values $z_t^{(k)}$. However, this can be misleading since it simply means that no value was proposed in the region where the likelihood is large. It does not guarantee that the target density $\widehat{f}_{t|t}\left(z_t|\widetilde{y}_{1:t}\right)$ has negligible mass there. A more detailed comparison between the rejection and the importance sampling methods in general can be found in Robert and Casella (2004), Section 3.3.3.

Of course, other proposal distributions than the prior can be used. As before, the disadvantage is that the sum in the acceptance probability should be computed. The idea of Pitt and Shephard (1999) to include explicitly the auxiliary index I was originally developed for this case. They proposed to generate a sample $((i_k, z_t^{(k)}), 1 \leq k \leq R)$ from the distribution $\tau(i)\rho\left(i, z_t\right)$ by generating first the index I_k according to a distribution $(\tau(i))$ and then the variable $Z_t^{(k)}$ according to a density $\rho\left(i_k, z_t\right)$ with $I_k = i_k$. After this, a sample of size N is selected with inclusion probabilities

$$\pi\left(i_k, z_t^{(k)}\right) \propto \frac{p\left(\widetilde{y}_t|z_t^{(k)}\right) p\left(z_t^{(k)}|z_{(t-p):(t-1)}^{(i_k)}\right)}{\tau(i_k)\rho\left(i_k, z_t^{(k)}\right)}.$$

In contrast to the rejection method, a promising idea here is to combine $\rho\left(i, z_t\right) = p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right)$ with unequal $\tau(i)$'s. For example, each $\tau(i)$ can be chosen to be proportional to $p\left(\widetilde{y}_t|m_i\right)$ where m_i is the mean or the median of $p\left(.|z_{(t-p):(t-1)}^{(i)}\right)$. If all $p\left(.|z_{(t-p):(t-1)}^{(i)}\right)$ have a small spread relative to the scale at which $p\left(\widetilde{y}_t|.\right)$ varies, then most $\pi\left(i_k, z_t^{(k)}\right)$ will be approximately equal, and therefore $R = N$ is sufficient.

2.3 Particle filtering recursion with $\widetilde{y}_t$ available

We have to distinguish two situations in the construction of the particle filtering recursion at time t: the case where $\widetilde{y}_t$ is available and the case where it is missing. The first case is considered in this section. In particular, the efforts to get an efficient and fast algorithm are explained. The particle filtering step with missing $\widetilde{y}_t$ is discussed in Section 2.4. This second case does not cause additional difficulties. In fact, the filtering step becomes simpler.

The crucial point in the implementation of the filtering recursion

(2.12) is an efficient scheme for sampling from the approximate density $\widehat{f}_{t|t}\left(z_t|\widetilde{y}_{1:t}\right)$ defined in (2.13). Two methods were presented to accomplish the general sampling task: the rejection sampling and the sampling importance resampling method (SIR).
Which method is better to apply? As briefly mentioned in the Subsection 2.2.2, the sampling importance resampling method can produce many ties in the final sample. Or it can be that the final sample contains no value in some regions although the target density $\widehat{f}_{t|t}\left(z_t|\widetilde{y}_{1:t}\right)$ has non-negligible mass there. For these reasons, we prefer the rejection method to sampling importance resampling. As the filtering algorithm should be as fast as possible, sampling with the rejection method should be efficient. The classical choice of the proposal density $\rho\left(z_t\right)$ is the prior (2.14). But this choice could be far from being optimal since the observation error distribution is heavy-tailed in the considered model (2.6) and (2.7). Consequently, the average acceptance probability could be small as remarked in the Subsection 2.2.1. It would be better to take directly $p\left(\widetilde{y}_t|z_t\right)$ as proposal density $\rho\left(z_t\right)$. But this choice is not free of difficulties either, since the evaluation of the acceptance probability requires the computation of the sum over i, see (2.15). The way out is given by the application of the rejection method with the auxiliary index. It also gives directly the used mixture indices in (2.13) which we need in order to return the sample $(z^{(l)}_{(t-p+1):t})$. In addition, the distribution $(\tau(i))$ and the densities $\rho\left(i, z_t\right)$ can be chosen in the optimal way described before. But unfortunately, the construction of the densities $\rho\left(i, z_t\right)$ is not straightforward. In fact, as a consequence of Lemma 2.1, each proposal $\rho\left(i, z_t\right)$ should be a trade-off between a good approximation of the density (2.17) and a density from which it is easy to sample. The ideal $\rho\left(i, z_t\right)$ is given by (2.17) itself. But this choice is not possible since the integral in its denominator cannot be written in closed form for the error distribution of the considered model (2.6) and (2.7). Thus, the key idea for the construction of the densities $\rho\left(i, z_t\right)$ is as follows. The "problematic" density $p\left(\widetilde{y}_t|z_t\right)$ in (2.17) is approximated by a majorant which is a mixture of exponential functions with arguments given by constant, linear or quadratic polynomials in z_t. Then this majorant is multiplied by the second density $p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)$ of (2.17) to get a mixture of normal densities which satisfies the previously mentioned trade-off. There is still something unpleasant in this idea. If the majorant is found directly for $p\left(\widetilde{y}_t|z_t\right)$, it will depend on the observation $\widetilde{y}_t$. Thus, a new majorant should be computed in each filtering step. We

can avoid this easily thanks to the following lemma.

Lemma 2.2 *Consider the model (2.6) and (2.7) and let Z_t be a random variable with probability density function $\widehat{f}_{t|t}(z_t|\widetilde{y}_{1:t})$ given by (2.13). Then the random variable Z defined by*

$$Z = \frac{Z_t - \widetilde{y}_t}{\sigma_V}$$

has density

$$\widehat{f}(z|\widetilde{y}_{1:t}) \propto p_{VII}(m, c/\sigma_V, 0)(z) \cdot \sum_{i=1}^{N} \phi\left(\frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, 1\right)(z) \tag{2.18}$$

with $(z_{(t-p):(t-1)}^{(i)})$ a sample of $Z_{(t-p):(t-1)}|\widetilde{Y}_{1:(t-1)} = \widetilde{y}_{1:(t-1)}$, $i = 1, \ldots, N$. φ_1, ..., φ_p, σ_V, m and c are the parameters of the model (2.6) and (2.7).

Proof: The proof is straightforward. We have:

$$\begin{aligned}
\widehat{f}_{t|t}(z_t|\widetilde{y}_{1:t}) &\propto p(\widetilde{y}_t|z_t) \cdot \sum_{i=1}^{N} p\left(z_t|z_{(t-p):(t-1)}^{(i)}\right) \\
&\propto b_{m,c}(z_t - \widetilde{y}_t) \cdot \sum_{i=1}^{N} \phi_{\sigma_V}\left(z_t - \sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)}\right) \\
&\propto b_{m,c}\left(\frac{z_t - \widetilde{y}_t}{\sigma_V}\sigma_V\right) \cdot \sum_{i=1}^{N} \phi_{\sigma_V}\left(z_t - \widetilde{y}_t - \left(\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t\right)\right) \\
&\propto \frac{1}{\sigma_V} \cdot b_{m,c/\sigma_V}\left(\frac{z_t - \widetilde{y}_t}{\sigma_V}\right) \cdot \sum_{i=1}^{N} \frac{1}{\sigma_V} \cdot \phi_1\left(\frac{z_t - \widetilde{y}_t}{\sigma_V} - \frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}\right) \\
&\propto b_{m,c/\sigma_V}(z) \cdot \sum_{i=1}^{N} \phi_1\left(z - \frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}\right).
\end{aligned}$$

Thus, the density of Z is:

$$\widehat{f}(z|\widetilde{y}_{1:t}) \propto p_{VII}(m, c/\sigma_V, 0)(z) \cdot \sum_{i=1}^{N} \phi\left(\frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, 1\right)(z).$$

□

As we can see, the Pearson type VII density in (2.18) does not depend on $\widetilde{Y}_t$. Thus, an efficient procedure to implement the filtering recursion at time t given $\widetilde{y}_t$ and a sample $\left(z^{(i)}_{(t-p):(t-1)}\right)$, $i = 1, \ldots, N$, of $Z_{(t-p):(t-1)}|\widetilde{Y}_{1:(t-1)}$ is as follows:

- Generate a sample $(z^{(l)})$ from the density $\widehat{f}\,(z|\widetilde{y}_{1:t})$ defined in (2.18).
- Set $z^{(l)}_t = \widetilde{y}_t + \sigma_V \cdot z^{(l)}$ for $l = 1, \ldots, N$. Then $(z^{(l)}_t)$ is a sample from the density $\widehat{f}_{t|t}\,(z_t|\widetilde{y}_{1:t})$ (see (2.13)).
- Return the sample $(z^{(l)}_{(t-p+1):t})$ of $Z_{(t-p+1):t}|\widetilde{Y}_{1:t}$.

The previous discussion about sampling from $\widehat{f}_{t|t}\,(z_t|\widetilde{y}_{1:t})$ also applies to sampling from $\widehat{f}\,(z|\widetilde{y}_{1:t})$. Therefore, we choose the rejection method with an auxiliary index to generate the sample $(z^{(l)})$ from $\widehat{f}\,(z|\widetilde{y}_{1:t})$.

The rest of this section is organized as follows. First, the construction of efficient proposal densities $\rho\,(i, z)$ and of the distribution $(\tau(i))$ is explained. Then, the acceptance probability of a pair (i, z) proposed using the rejection method with the auxiliary index is computed. Last, the implementation of the filtering recursion at time t with $\widetilde{y}_t$ available is summarized.

2.3.1 Construction of proposal densities $\rho\,(i, z)$

We saw that each proposal density $\rho\,(i, z)$ should be a trade-off between a good proposal distribution for the density proportional to

$$p_{VII}\,(m, c/\sigma_V, 0)\,(z) \cdot \phi\left(\frac{\sum_{l=1}^{p} \varphi_l z^{(i)}_{t-l} - \widetilde{y}_t}{\sigma_V}, 1\right)(z)$$

and a distribution from which it is easy to sample. The key idea was to approximate the density $p_{VII}\,(m, c/\sigma_V, 0)\,(z)$ by a clever majorant. Actually, it is easier to approximate the natural logarithm of this density up to some terms. Since

$$\log\left(p_{VII}\,(m, c/\sigma_V, 0)\,(z)\right) = \log\left(\frac{\Gamma(m)\sigma_V}{\sqrt{\pi}c\Gamma(m-0.5)}\right) - m\log\left[1 + \left(\frac{z}{c/\sigma_V}\right)^2\right],$$

the unique term on the right side which depends on z is given by $-\log\left[1 + \left(\frac{z}{c/\sigma_V}\right)^2\right]$. It suffices to seek the majorant for it.

Definition 2.2 *For a given integer value K, we choose parameters α_k, β_k, γ_k and d_k such that*

$$-\log\left[1+\left(\frac{z}{c/\sigma_V}\right)^2\right] \leq \sum_{k=1}^{K}\left(\alpha_k+\beta_k z+\gamma_k z^2\right)\mathbb{I}_{\{z\in B_k\}} \qquad \forall z\in\mathbb{R} \tag{2.19}$$

with $B_k := (d_{k-1}, d_k]$, $-\infty =: d_0 < d_1 < \cdots < d_K := \infty$. The parameters α_k, β_k, γ_k and d_k are not allowed to depend on z. log *denotes the natural logarithm.*

Remark 2.5 *Some comments about the previous definition.*

- *The function*

  $$g(z) := -\log\left[1+\left(\frac{z}{c/\sigma_V}\right)^2\right]$$

  is negative for all z in $\mathbb{R}$, symmetric about 0 *and it has a maximum in* 0.
- *In words, the majorant consists of a mixture of quadratic polynomials on the intervals B_k, $k=1,\dots,K$. Note that*

  $$\bigcup_{k=1}^{K} B_k = \mathbb{R} \qquad \text{and} \qquad B_k\cap B_{k'}=\emptyset \quad \text{for } k\neq k'.$$

- *The reason why we need an approximation by a majorant will be clear when we compute the optimal distribution $(\tau(i))$, see Subsection 2.3.2. Moreover, the majorant should be constructed such that it is as close as possible to the function $g(z)$. This is important to have a high acceptance probability in the rejection method with the auxiliary index, see Subsection 2.3.3.*

We postpone the construction of the majorant. At the moment, it is more interesting to show how we use it.

Lemma 2.3 *Let the proposal densities $\rho(i,z)$ be chosen as*

$$\rho(i,z)=\sum_{k=1}^{K}\frac{R_{k,i}M_{k,i}}{\sum_{l=1}^{K}R_{l,i}M_{l,i}}\;\frac{\exp\left[-\frac{1}{2}\left(\frac{z-\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)^2\right]}{M_{k,i}}\;\mathbb{I}_{\{z\in B_k\}}$$

with

$$
\begin{aligned}
\overline{\sigma}_k &:= \sqrt{\frac{1}{1-2m\gamma_k}}, \\
\mu_i &:= \frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, \\
\overline{\mu}_{k,i} &:= \overline{\sigma}_k^2 \;\left(\mu_i + m\beta_k\right), \\
R_{k,i} &:= \exp\left[\frac{1}{2}\left(\frac{\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)^2 - \frac{1}{2}\mu_i^2 + m\alpha_k\right], \\
M_{k,i} &:= \sqrt{2\pi}\;\overline{\sigma}_k \left[\Phi\left(\frac{d_k - \overline{\mu}_{k,i}}{\overline{\sigma}_k}\right) - \Phi\left(\frac{d_{k-1} - \overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)\right].
\end{aligned}
$$

The parameters α_k, β_k, γ_k and d_k are as in Definition 2.2; σ_V, m and c are the parameters of the observation error distribution in the considered model (2.6) and (2.7). Finally, $\Phi\left(x\right)$ is the cumulative $\mathcal{N}(0,1)$ distribution function.

Then

$$
p_{VII}\left(m, c/\sigma_V, 0\right)(z)\cdot\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, 1\right)(z) \le k_1\cdot\Big(\sum_{l=1}^{K} R_{l,i}M_{l,i}\Big)\cdot\rho\left(i,z\right) \tag{2.20}
$$

with

$$
k_1 := \frac{\Gamma(m)\;\sigma_V}{\sqrt{2}\;\pi\;c\;\Gamma(m-0.5)}.
$$

Remark 2.6 *Some comments about the lemma before we prove it.*

- *The μ_i's are the expected values of the normal densities in the target density (2.18). Note the form: they are a normalized difference between the prediction value of Z_t according to an AR(p) process (the state equation (2.6)) and the observed value $\widetilde{y}_t$. If $\widetilde{y}_t$ is not an outlier, the absolute values of the μ_i's are small. This is true in most cases.*
- *The term*

  $$
  \frac{\exp\left[-\frac{1}{2}\left(\frac{z-\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)^2\right]}{M_{k,i}}
  $$

  is the density of the truncated $\mathcal{N}(\overline{\mu}_{k,i}, \overline{\sigma}_k^2)$ distribution on B_k. Therefore, each proposal $\rho\left(i,z\right)$ is a mixture of truncated normal *densities with weights $\omega_i(k)$ given by*

$$\omega_i(k) := \frac{R_{k,i} M_{k,i}}{\sum_{l=1}^{K} R_{l,i} M_{l,i}} \ .$$

The proposals $\rho\,(i,z)$ satisfy the previously mentioned trade-off provided that the majorant is a good approximation of the function $g(z)$.

- *The sampling of Z from $\rho\,(i,z)$ with known $I = i$ can be carried out by the following two-step procedure. First, the index K^* is generated according to the weights distribution $(\omega_i(k))$ using the inversion method. This requires the evaluation of the weights partial sums. Then, the variable Z is sampled from the truncated $\mathcal{N}(\overline{\mu}_{k^*,i}, \overline{\sigma}^2_{k^*})$ density on B_{k^*} with $K^* = k^*$ using again the inversion method. Explicitly, define*

  $$b_{k^*} := \Phi\left(\frac{d_{k^*-1} - \overline{\mu}_{k^*,i}}{\overline{\sigma}_{k^*}}\right) \qquad \text{and} \qquad c_{k^*} := \Phi\left(\frac{d_{k^*} - \overline{\mu}_{k^*,i}}{\overline{\sigma}_{k^*}}\right).$$

  If U is uniformly distributed on $[0,1]$, then

  $$U^* := b_{k^*} + (c_{k^*} - b_{k^*}) \cdot U$$

  is uniformly distributed on $[b_{k^*}, c_{k^*}]$. Thus

  $$Z := \overline{\mu}_{k^*,i} + \overline{\sigma}_{k^*} \cdot \Phi^{-1}\left(U^*\right)$$

  is distributed according to the truncated $\mathcal{N}(\overline{\mu}_{k^*,i}, \overline{\sigma}^2_{k^*})$ distribution on B_{k^*}.

  More details on the inversion method can be found in Ripley (1987), Sections 3.2 and 3.3. Geweke (1991) proposed a more efficient procedure for the sampling from a truncated normal distribution.

- *Inequality (2.20) will be useful later to find the optimal distribution $(\tau(i))$, see Subsection 2.3.2.*

  The inequality can be illustrated also in another way. Let $q\,(z)$ be a majorant of the density $p_{VII}\,(m, c/\sigma_V, 0)\,(z)$ and not of its logarithm (up to some terms) as it is in Definition 2.2. It follows that

  $$p_{VII}\,(m, c/\sigma_V, 0)\,(z) \cdot \phi\left(\frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, 1\right)(z)$$

  $$\leq q\,(z) \cdot \phi\left(\frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, 1\right)(z)$$

$$=\int q\left(z\right)\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)}-\widetilde{y}_t}{\sigma_V},1\right)(z)\,dz\cdot\frac{q\left(z\right)\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)}-\widetilde{y}_t}{\sigma_V},1\right)(z)}{\int q\left(z\right)\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)}-\widetilde{y}_t}{\sigma_V},1\right)(z)\,dz}. \tag{2.21}$$

The normalising constant

$$\int q\left(z\right)\phi_1\left(z-\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)}-\widetilde{y}_t}{\sigma_V}\right)dz$$

has an interesting feature: it is the convolution of the two functions $q\left(z\right)$ *and* $\phi_1\left(z\right)$ *computed in the point* $\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)}-\widetilde{y}_t}{\sigma_V}$*. This property will be useful to implement the smoothing recursion using the second method, see Subsection 3.4.1.*
Note that in Lemma 2.3, we find explicitly the expression (2.21) for a specific choice of $\log(q\left(z\right))$ *and we denoted it by*

$$k_1\cdot\left(\sum_{l=1}^{K}R_{l,i}M_{l,i}\right)\cdot\rho\left(i,z\right).$$

Proof of Lemma 2.3:
This is a constructive proof. The proposals $\rho\left(i,z\right)$ are constructed using the definition of the majorant (Definition 2.2) and completing the square in the involved exponential terms. The inequality (2.20) follows directly from the construction. Note that we generalize the proof a little bit by setting the variance of the involved normal density equal to λ. In this way, we can adapt the proof easily to the smoothing case. But when we use the notations introduced in this lemma, we assume $\lambda=1$.

First, we find using the definition of the majorant:

$$\log\left[p_{VII}\left(m,c/\sigma_V,0\right)(z)\cdot\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)}-\widetilde{y}_t}{\sigma_V},\sqrt{\lambda}\right)(z)\right]$$

$$=\log\left\{\frac{\Gamma(m)}{\sqrt{\pi}\ \frac{c}{\sigma_V}\ \Gamma(m-0.5)}\cdot\frac{1}{\left[1+\left(\frac{z}{c/\sigma_V}\right)^2\right]^m}\right\}+$$

$$+\log\left\{\frac{1}{\sqrt{2\pi\lambda}}\exp\left[-\frac{1}{2\lambda}\left(z-\mu_i\right)^2\right]\right\}$$

$$\begin{aligned}
&= \log\left(k_1\right) - m\ \log\left[1 + \left(\frac{z}{c/\sigma_V}\right)^2\right] - \frac{1}{2\lambda}\left(z-\mu_i\right)^2 \qquad (2.22)\\
&\leq \log\left(k_1\right) + \left[\sum_{k=1}^{K}\left(m\alpha_k + m\beta_k z + m\gamma_k z^2\right)\mathbb{I}_{\{z\in B_k\}}\right] - \frac{1}{2\lambda}\left(z-\mu_i\right)^2\\
&= \log\left(k_1\right) + \sum_{k=1}^{K}\left[m\alpha_k + m\beta_k z + m\gamma_k z^2 - \frac{1}{2\lambda}\left(z-\mu_i\right)^2\right]\mathbb{I}_{\{z\in B_k\}}.
\end{aligned}$$

It follows:

$$\begin{aligned}
&p_{VII}\left(m, c/\sigma_V, 0\right)(z)\cdot\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, \sqrt{\lambda}\right)(z)\\
&\leq \exp\left\{\log\left(k_1\right) + \sum_{k=1}^{K}\left[m\alpha_k + m\beta_k z + m\gamma_k z^2 - \frac{1}{2\lambda}\left(z-\mu_i\right)^2\right]\mathbb{I}_{\{z\in B_k\}}\right\}\\
&= k_1\sum_{k=1}^{K}\exp\left[m\alpha_k + m\beta_k z + m\gamma_k z^2 - \frac{1}{2\lambda}\left(z-\mu_i\right)^2\right]\mathbb{I}_{\{z\in B_k\}}.
\end{aligned}$$

Now, we get completing the square in the exponents:

$$\begin{aligned}
&m\alpha_k + m\beta_k z + m\gamma_k z^2 - \frac{1}{2\lambda}\left(z-\mu_i\right)^2\\
&= -\frac{1}{2}\left[\left(\frac{1}{\lambda} - 2m\gamma_k\right)z^2 - 2\left(m\beta_k + \frac{\mu_i}{\lambda}\right)z + \frac{\mu_i^2}{\lambda} - 2m\alpha_k\right]\\
&= -\frac{1}{2}\left(\frac{1}{\lambda} - 2m\gamma_k\right)\left[z^2 - 2\left(\frac{1}{\lambda} - 2m\gamma_k\right)^{-1}\left(m\beta_k + \frac{\mu_i}{\lambda}\right)z\right] -\\
&\quad -\frac{1}{2}\left(\frac{\mu_i^2}{\lambda} - 2m\alpha_k\right)\\
&= -\frac{1}{2}\left(\frac{1}{\lambda} - 2m\gamma_k\right)\left[z - \left(\frac{1}{\lambda} - 2m\gamma_k\right)^{-1}\left(m\beta_k + \frac{\mu_i}{\lambda}\right)\right]^2 +\\
&\quad +\frac{1}{2}\left(\frac{1}{\lambda} - 2m\gamma_k\right)^{-1}\left(m\beta_k + \frac{\mu_i}{\lambda}\right)^2 - \frac{\mu_i^2}{2\lambda} + m\alpha_k\\
&= -\frac{1}{2}\,\frac{1}{\overline{\sigma}_k^2}\,\left(z - \overline{\mu}_{k,i}\right)^2 + \frac{1}{2}\left(\frac{\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)^2 - \frac{\mu_i^2}{2\lambda} + m\alpha_k\\
&= -\frac{1}{2}\left(\frac{z - \overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)^2 + \log\left(R_{k,i}\right). \qquad (2.23)
\end{aligned}$$

Altogether:

$$
\begin{aligned}
&p_{VII}\left(m, c/\sigma_V, 0\right)(z) \cdot \phi\left(\frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, \sqrt{\lambda}\right)(z) \\
&\leq k_1 \sum_{k=1}^{K} \exp\left[m\alpha_k + m\beta_k z + m\gamma_k z^2 - \frac{1}{2\lambda}\left(z-\mu_i\right)^2\right] \mathbb{I}_{\{z \in B_k\}} \\
&= k_1 \sum_{k=1}^{K} R_{k,i}\ \exp\left[-\frac{1}{2}\left(\frac{z-\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)^2\right] \mathbb{I}_{\{z \in B_k\}}.
\end{aligned}
$$

Finally, we normalize the right hand side:

$$
\begin{aligned}
&p_{VII}\left(m, c/\sigma_V, 0\right)(z) \cdot \phi\left(\frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, \sqrt{\lambda}\right)(z) \\
&\leq k_1 \cdot \left(\sum_{l=1}^{K} R_{l,i} M_{l,i}\right) \cdot \sum_{k=1}^{K} \frac{R_{k,i} M_{k,i}}{\sum_{l=1}^{K} R_{l,i} M_{l,i}} \, \frac{\exp\left[-\frac{1}{2}\left(\frac{z-\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)^2\right]}{M_{k,i}} \mathbb{I}_{\{z \in B_k\}} \\
&= k_1 \cdot \left(\sum_{l=1}^{K} R_{l,i} M_{l,i}\right) \cdot \rho\left(i, z\right).
\end{aligned}
$$

The density feature of the proposals $\rho\left(i, z\right)$ follows easily from their definition, see also Remark 2.6. □

2.3.2 Construction of the distribution $(\tau(i))$

The second point to discuss in the application of the rejection method with the auxiliary index is the construction of the distribution $(\tau(i))$ once the proposal densities $\rho\left(i, z\right)$ are given. Lemma 2.1 gives the optimal choice. We find:

Lemma 2.4 *Let the densities $\rho\left(i, z\right)$ be defined as in Lemma 2.3. Then the optimal distribution $(\tau(i))$ is given by*

$$
\tau(i) = \frac{\sum_{l=1}^{K} R_{l,i} M_{l,i}}{\sum_{i=1}^{N} \sum_{l=1}^{K} R_{l,i} M_{l,i}}.
$$

Proof: The lemma follows easily using Lemma 2.1 and the inequality (2.20). Note that the target density is given by $\widehat{f}\,(z|\widetilde{y}_{1:t})$, see (2.18). From Lemma 2.1, the optimal $\tau(i)$'s are given by

$$\tau(i) \ \propto M_i$$

with

$$M_i \geq \sup_z \frac{p_{VII}\,(m, c/\sigma_V, 0)\,(z)\,\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, 1\right)(z)}{\rho\,(i,z)}.$$

Using the inequality (2.20), we find

$$\frac{p_{VII}\,(m, c/\sigma_V, 0)\,(z)\,\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, 1\right)(z)}{\rho\,(i,z)} \leq k_1 \sum_{l=1}^{K} R_{l,i} M_{l,i}.$$

M_i can be set equal to the right hand side, since the latter is independent of z. The lemma follows since k_1 is independent of the past sample $(z_{(t-p):(t-1)}^{(i)})$. □

2.3.3 Acceptance probability of the proposed pair (i, z)

In the last Subsections 2.3.1 and 2.3.2, we have constructed useful proposal densities $\rho\,(i,z)$ and the optimal distribution $(\tau(i))$. Thus, we are able to generate a pair (I, Z) according to the distribution $\tau(i)\rho\,(i,z)$ by first sampling the auxiliary index I according to $(\tau(i))$ and then the variable Z according to the density $\rho\,(i,z)$ with $I = i$. The next step consists of evaluating the acceptance probability of the proposed pair (i,z).

Lemma 2.5 *Let the densities $\rho\,(i,z)$ and the distribution $(\tau(i))$ be defined as in Lemmas 2.3 and 2.4, respectively.*
Then the acceptance probability of the pair (i,z) generated from the distribution $\tau(i)\rho\,(i,z)$ is

$$\pi\,(i,z) = \exp\left\{-m\left[\log\left(1+\left(\frac{z}{c/\sigma_V}\right)^2\right) + \alpha_{k*} + \beta_{k*}z + \gamma_{k*}z^2\right]\right\}$$

where B_{k*} is the subset containing z, see Definition 2.2.

Remark 2.7 *The resulting acceptance probability in Lemma 2.5 is not a surprise. It is a direct consequence of the majorant Definition 2.2 and of the rejection methodology. In fact, approximation (2.19) permits the construction of useful proposal densities $\rho(i,z)$ starting from the ideal ones. Then (2.19) also affects the acceptance step: it measures how good the proposal densities $\rho(i,z)$ are in comparison with the ideal ones (which imply $\pi(i,z) = 1$). See also how the acceptance probabilities (2.15) and (2.16) are constructed. Therefore, the majorant should approximate the function $g(z)$ well to have a high acceptance probability. This feature will help us to construct the majorant in Subsection 2.3.4.*

Proof: The lemma follows by taking the formula of the acceptance probability in the rejection method with the auxiliary index and evaluating it with the proposal densities $\rho(i,z)$ and the distribution $(\tau(i))$ (see their definitions in the Lemmas 2.3 and 2.4). Note that the target density is given by $\widehat{f}(z|\widetilde{y}_{1:t})$, see (2.18). In addition, we generalize again the proof by setting the variance of the involved normal density equal to λ. In this way, we can adapt the proof easily to the smoothing case. But when we use the introduced notations, we suppose $\lambda = 1$.

The acceptance probability is given similarly to (2.16) by

$$\pi(i,z) = \frac{p_{VII}(m, c/\sigma_V, 0)(z)\,\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, \sqrt{\lambda}\right)(z)}{\tau(i)\ \rho(i,z)\ \ M}$$

with

$$M \geq \sup_{i,z} \frac{p_{VII}(m, c/\sigma_V, 0)(z)\,\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, \sqrt{\lambda}\right)(z)}{\tau(i)\ \rho(i,z)}.$$

First, we calculate the supremum term and we define M. This can be achieved using the inequality (2.20) and the definition of the distribution $(\tau(i))$:

$$\begin{aligned}\frac{p_{VII}(m, c/\sigma_V, 0)(z)\,\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}, \sqrt{\lambda}\right)(z)}{\tau(i)\ \rho(i,z)} &\leq \frac{k_1\ \sum_{l=1}^{K} R_{l,i} M_{l,i}}{\tau(i)} \\ &= k_1 \sum_{i=1}^{N}\sum_{l=1}^{K} R_{l,i} M_{l,i}.\end{aligned}$$

The last expression does not depend neither on z nor on the past sample $(z^{(i)}_{(t-p):(t-1)})$. Therefore, M can be set equal to it. Then it follows for the acceptance probability:

$$\begin{aligned}
\pi\left(i,z\right) &= \frac{p_{VII}\left(m, c/\sigma_V, 0\right)(z)\,\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z^{(i)}_{t-l}-\widetilde{y}_t}{\sigma_V}, \sqrt{\lambda}\right)(z)}{k_1\ \left(\sum_{i=1}^{N}\sum_{l=1}^{K} R_{l,i}M_{l,i}\right)\ \tau(i)\ \rho\left(i,z\right)} \\
&= \frac{p_{VII}\left(m, c/\sigma_V, 0\right)(z)\,\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z^{(i)}_{t-l}-\widetilde{y}_t}{\sigma_V}, \sqrt{\lambda}\right)(z)}{k_1\ \left(\sum_{l=1}^{K} R_{l,i}M_{l,i}\right)\ \rho\left(i,z\right)} \\
&= \frac{p_{VII}\left(m, c/\sigma_V, 0\right)(z)\,\phi\left(\frac{\sum_{l=1}^{p}\varphi_l z^{(i)}_{t-l}-\widetilde{y}_t}{\sigma_V}, \sqrt{\lambda}\right)(z)}{k_1\ R_{k*,i}\ \exp\left[-\frac{1}{2}\left(\frac{z-\overline{\mu}_{k*,i}}{\sigma_{k*}}\right)^2\right]} \\
&= \frac{k_1\ \exp\left\{-m\log\left[1+\left(\frac{z}{c/\sigma_V}\right)^2\right]-\frac{1}{2\lambda}\left(z-\mu_i\right)^2\right\}}{k_1\ \exp\left[m\alpha_{k*}+m\beta_{k*}z+m\gamma_{k*}z^2-\frac{1}{2\lambda}\left(z-\mu_i\right)^2\right]} \\
&= \frac{\exp\left\{-m\log\left[1+\left(\frac{z}{c/\sigma_V}\right)^2\right]\right\}}{\exp\left(m\alpha_{k*}+m\beta_{k*}z+m\gamma_{k*}z^2\right)}
\end{aligned} \tag{2.24}$$

using successively the definitions of M, $\tau(i)$, $\rho\left(i,z\right)$ and (2.22), (2.23). Note that we know which term of the mixture $\rho\left(i,z\right)$ we should take, since z is given. Thus, the lemma follows. □

2.3.4 Construction of majorants

In Subsection 2.3.1, we have postponed the construction of the majorant in Definition 2.2. We have now all elements to construct it.

We note in Remark 2.7 that the majorant should be as close as possible to the function $g(z) = -\log\left[1+\left(\frac{z}{c/\sigma_V}\right)^2\right]$. In this way, the proposed pairs (i,z) have a high acceptance probability. Above all, the majorant should approximate $g(z)$ well in the regions where the target density $\widehat{f}\left(z|\widetilde{y}_{1:t}\right)$ has relevant mass since the proposed particles $(z^{(l)})$

come from these regions. How can we find these regions? We see from the definition of $\widehat{f}\,(z|\widetilde{y}_{1:t})$ in (2.18) that the Pearson type VII density has main mass around zero whereas the normal densities are concentrated around their expected values

$$\mu_i = \frac{\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)} - \widetilde{y}_t}{\sigma_V}.$$

Therefore, the target density $\widehat{f}\,(z|\widetilde{y}_{1:t})$ will have relevant mass in some places between 0 and the regions where μ_i lie. A bimodal $\widehat{f}\,(z|\widetilde{y}_{1:t})$ cannot be excluded, as Figure 2.2 shows. In this Figure, the resulting target density $\widehat{f}\,(z|\widetilde{y}_{1:t})$ is shown for different values of $\widetilde{y}_t$ given 10 fixed values of $\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)}$. The comparison is made clearer by rescaling the resulting target densities such that they have all the same maximal value. In addition, the fixed values $\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)}$ are given as vertical bars on the horizontal axis. The other parameters have values $\sigma_V = 2$, $m = 2$ and $c = 1$.

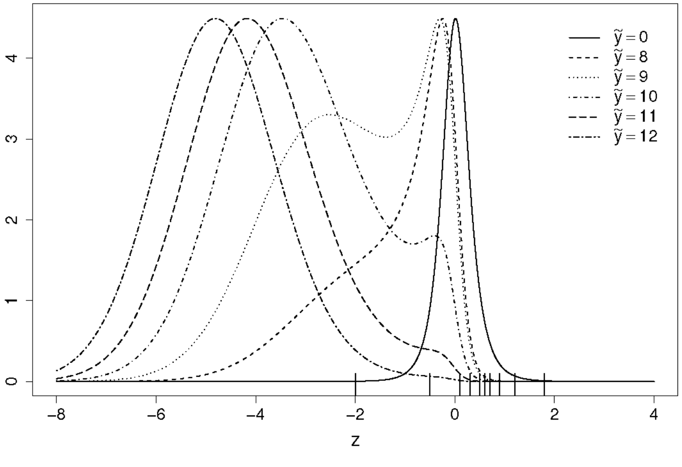

Figure 2.2: *Rescaled target density $\widehat{f}\,(z|\widetilde{y}_{1:t})$ for different values of $\widetilde{y}_t$ given 10 fixed values of $\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)}$ which are shown by vertical bars on the horizontal axis. The other parameters are $\sigma_V = 2$, $m = 2$ and $c = 1$.*

On the other hand, the majorant should be easy and fast to compute since it affects the construction of both the proposal densities $\rho(i,z)$ and the distribution $(\tau(i))$. Therefore, the number K of the mixture subsets should be small and the mixture terms should be preferably constant or linear functions. In addition, we want to compute the majorant as rarely as possible in the whole filtering recursion to speed up the algorithm. For this reason, we have reformulated the sampling at time step t using Lemma 2.2.
How can we take into account all these features? We consider three majorants, which we call the default, the lower and the upper majorant. All three majorants are mixtures with 7 terms (i.e. $K = 7$). We choose one majorant at the beginning of each filtering step and we use it to sample all particles $(z^{(l)})$ for this step. The default majorant is constructed to approximate very well the region around zero. It is computed only once at the beginning of the filtering recursion since its components depend only on the parameters σ_V and c (scale parameters of the considered error distributions). The default majorant is selected in the filtering steps where the μ_i's are around zero. Thus, it will be the most frequently chosen majorant since the μ_i's are typically around zero (see Remark 2.6). On the other hand, if the median of the μ_i's exceeds a lower or an upper bound, then we should construct a majorant which approximates well the regions both around zero and around the median. This idea leads to the definition of the lower and the upper majorant. Some components of them will depend on the median and thus, if the lower or the upper majorant is chosen in one filtering step, we have to compute first these components. The lower or the upper majorants are selected only in few filtering steps. Actually, in steps where the observed values $(\widetilde{y}_t)$ are outliers and therefore the μ_i's are big (in absolute value). The choice of the lower or the upper majorant guarantees that the sampling remains efficient also in such problematic cases (see the previous discussion).

Definition 2.3 *Let the μ_i's be as in Lemma 2.3 and let δ be a strictly positive real number. Define*

$$\mu := med_i\,(\mu_i)\,,$$

$$l_{(-)} := g(\mu - \delta) = -\log\left[1 + \left(\frac{\mu - \delta}{c/\sigma_V}\right)^2\right],$$

$$l_{(+)} := g(\mu+\delta) = -\log\left[1+\left(\frac{\mu+\delta}{c/\sigma_V}\right)^2\right].$$

Then the parameters to define the default majorant are given in Table 2.1, the parameters for the lower majorant in Table 2.2 and the parameters for the upper majorant in Table 2.3. Note that lower and upper majorants can be defined only if $\mu < -\sqrt{3}\ c/\sigma_V - \delta$ and $\mu > \sqrt{3}\ c/\sigma_V + \delta$, respectively.

k	d_{k-1}	d_k	α_k	β_k	γ_k
1	$-\infty$	$-5\ c/\sigma_V$	$-\log(26)$	0	0
2	$-5\ c/\sigma_V$	$-\sqrt{3}\ c/\sigma_V$	$\frac{\sqrt{3}\log(6.5)}{5-\sqrt{3}} - \log(4)$	$\frac{\log(6.5)}{\left(5-\sqrt{3}\right)\ c/\sigma_V}$	0
3	$-\sqrt{3}\ c/\sigma_V$	$-c/\sigma_V$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\sigma_V}$	0
4	$-c/\sigma_V$	c/σ_V	0	0	$-\frac{\log(2)}{(c/\sigma_V)^2}$
5	c/σ_V	$\sqrt{3}\ c/\sigma_V$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$-\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\sigma_V}$	0
6	$\sqrt{3}\ c/\sigma_V$	$5\ c/\sigma_V$	$\frac{\sqrt{3}\log(6.5)}{5-\sqrt{3}} - \log(4)$	$-\frac{\log(6.5)}{\left(5-\sqrt{3}\right)\ c/\sigma_V}$	0
7	$5\ c/\sigma_V$	∞	$-\log(26)$	0	0

Table 2.1: *Parameters to define the default majorant in the filtering recursion.*

We select

- *the default majorant if* $-\sqrt{3}\ c/\sigma_V - \delta \le \mu \le \sqrt{3}\ c/\sigma_V + \delta$,
- *the lower majorant if* $-\sqrt{3}\ c/\sigma_V - \delta > \mu$,
- *the upper majorant if* $\mu > \sqrt{3}\ c/\sigma_V + \delta$.

Remark 2.8 *Some comments about the definition.*

- *The majorants depend on few parameters: c/σ_V, μ and δ. The latter two parameters are used only in the construction of lower and upper majorants. In addition, the polynomials to define the majorants are chosen as simple as possible. Therefore, we choose a quadratic polynomial (a parabola) only for the approximations around zero. On the first and the last subsets, it is sufficient an approximation by a constant. On the other subsets, the majorants are given by a secant (linear polynomial) through the begin and the*

k	d_{k-1}	d_k	α_k	β_k	γ_k
1	$-\infty$	$\mu-\delta$	$l_{(-)}$	0	0
2	$\mu-\delta$	$\mu+\delta$	$l_{(-)}-\beta_2\ (\mu-\delta)$	$\frac{l_{(+)}-l_{(-)}}{2\ \delta}$	0
3	$\mu+\delta$	$-\sqrt{3}\ c/\sigma_V$	$l_{(+)}-\beta_3\ (\mu+\delta)$	$\frac{\log(4)+l_{(+)}}{\mu+\delta+\sqrt{3}\ c/\sigma_V}$	0
4	$-\sqrt{3}\ c/\sigma_V$	$-c/\sigma_V$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\sigma_V}$	0
5	$-c/\sigma_V$	c/σ_V	0	0	$-\frac{\log(2)}{(c/\sigma_V)^2}$
6	c/σ_V	$\sqrt{3}\ c/\sigma_V$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$-\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\sigma_V}$	0
7	$\sqrt{3}\ c/\sigma_V$	∞	$-\log(4)$	0	0

Table 2.2: *Parameters to define the lower majorant in the filtering recursion. We should have $\mu < -\sqrt{3}\ c/\sigma_V - \delta$.*

k	d_{k-1}	d_k	α_k	β_k	γ_k
1	$-\infty$	$-\sqrt{3}\ c/\sigma_V$	$-\log(4)$	0	0
2	$-\sqrt{3}\ c/\sigma_V$	$-c/\sigma_V$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\sigma_V}$	0
3	$-c/\sigma_V$	c/σ_V	0	0	$-\frac{\log(2)}{(c/\sigma_V)^2}$
4	c/σ_V	$\sqrt{3}\ c/\sigma_V$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$-\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\sigma_V}$	0
5	$\sqrt{3}\ c/\sigma_V$	$\mu-\delta$	$-\log(4)-\beta_5\ \sqrt{3}\ c/\sigma_V$	$\frac{l_{(-)}+\log(4)}{\mu-\delta-\sqrt{3}\ c/\sigma_V}$	0
6	$\mu-\delta$	$\mu+\delta$	$l_{(-)}-\beta_6\ (\mu-\delta)$	$\frac{l_{(+)}-l_{(-)}}{2\ \delta}$	0
7	$\mu+\delta$	∞	$l_{(+)}$	0	0

Table 2.3: *Parameters to define the upper majorant in the filtering recursion. We should have $\mu > \sqrt{3}\ c/\sigma_V + \delta$.*

end point of the subsets. In addition, since the region around zero should be approximated well by all majorants, the corresponding subsets and polynomials are the same. Moreover, note that all γ_k's are zero or strictly negative. Thus, the defined variances $\overline{\sigma}_k^2$ are always positive (see their definition in Lemma 2.3).

- *With δ we take into account the fact that the $\mathcal{N}(\mu_i, 1)$ densities in (2.18) have their main mass in the regions around the corre-*

sponding expected values μ_i. The interval $[\mu-\delta, \mu+\delta]$ covers such regions and the lower and upper majorants are constructed to give also a good approximation on this interval. We take the median of the μ_i's to define the interval since we want to avoid a direct dependence on the sample $(z^{(i)}_{(t-p):(t-1)})$. The default value of δ is two.

- *If μ is less than the bound $-\sqrt{3}\ c/\sigma_V-\delta$, we put aside the default majorant and we choose the lower one. In most cases, this lower bound is still in a region where the approximation given by the default majorant would be good. In this way, we have a smooth transition from the default majorant to the lower one.*
  The cases with a very small value of c/σ_V are an exception. In fact, the approximation given by the default majorant becomes bad away from $|5c/\sigma_V|$ which is small in these cases. But the lower majorant is selected first by $\mu<-\sqrt{3}c/\sigma_V-\delta\approx-\delta=-2$. Then, the transition between the default and the lower majorant is not so smooth as before and, consequently, the sampling of $(z^{(l)})$ generates more rejections. The efficiency loss is not huge. For small values of c/σ_V, the state errors mask the heavy-tailed observation errors and, consequently, the observed values $(\widetilde{Y}_t)$ follow the pattern of the state variables (Z_t). Thus, the μ_i's are near zero and the default majorant is chosen.
  Alternatively, one could introduce a second lower majorant to cover the cases with $-\sqrt{3}c/\sigma_V-\delta\le\mu<-\sqrt{3}c/\sigma_V$. Then a high efficiency is also attained for the situations with a small c/σ_V value. Since the efficiency loss is low, we do not follow this idea. Of course, an equivalent discussion is valid for the upper majorant.

- *Other problematic situations arise when about 50% of the μ_i's is less than $-5c/\sigma_V$ and the other 50% is greater than $5c/\sigma_V$. The default majorant is chosen but it is not a good approximation in such situations.*
  These cases may happen for example in presence of an observation outlier: the target density has non-negligible mass in two distinct regions. Actually, we have never got problems in these situations (or perhaps we have never encountered these cases in simulated or real examples).

- *As said, the lower and the upper majorants should approximate well the function $g(z)$ on the interval $[\mu-\delta,\mu+\delta]$ and not only around zero. On the other hand, we would like to retain the same* *number of mixture terms as in the default case to avoid difficulties in the computer implementation. Thus, we simplify lower and upper majorants on the side where the median does not lie. The resulting majorants are no more symmetric about zero as it was for the default one.*

- *The default and the upper majorants are shown in Figure 2.3 for the special case $c/\sigma_V = 1$, $\mu = 8$ and $\delta = 2$. We see that the approximations are very good "where it is needed" (around zero and μ).*

Lemma 2.6 *The three majorants in Definition 2.3 satisfy Definition 2.2. I.e., they fulfil the inequality*

$$g(z) = -\log\left[1 + \left(\frac{z}{c/\sigma_V}\right)^2\right] \leq \sum_{k=1}^{K} \left(\alpha_k + \beta_k z + \gamma_k z^2\right) \mathbb{I}_{\{z \in B_k\}}, \ \forall z \in \mathbb{R}.$$

In addition, these majorants are continuous functions in z.

Proof: The proof of the inequality is very instructive. In fact, it illustrates how the majorants are constructed.
Without loss of generality, we consider

$$f(x) := -\log\left(1 + x^2\right)$$

(set $x = z/\left(c/\sigma_V\right)$). Moreover, the three cases are proved together.
The proof is easy on the first and on the 7^{th} subset of the mixtures since the chosen polynomials are constant on these subsets and the constants are the value of $f(x)$ in the end point of B_1 and in the start point of B_7, respectively.
The first two derivatives of $f(x)$ are given by

$$\begin{aligned} f'(x) &= -\frac{2x}{1+x^2}, \\ f''(x) &= -\frac{2(1+x^2) - 2x\ 2x}{\left(1+x^2\right)^2} = 2\ \frac{x^2-1}{\left(1+x^2\right)^2}. \end{aligned}$$

Then $f''(x)$ is strictly positive on the subset $\mathbb{R} \setminus [-1, 1]$. It follows that $f(x)$ is strictly convex on this subset and therefore all secants between two points $(b_1, f(b_1))$, $(b_2, f(b_2))$, with $b_1 < b_2 < -1$ or $1 < b_1 < b_2$, are automatically majorants. This proves the inequality on the mixture subsets approximated by linear polynomials.

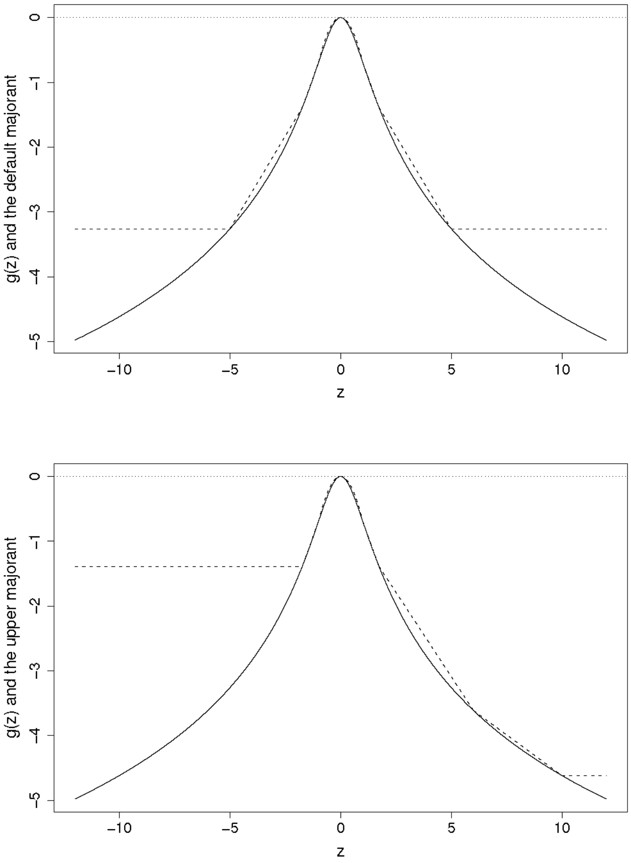

Figure 2.3: *The default and the upper majorants for the case* $c/\sigma_V = 1$*,* $\mu = 8$ *and* $\delta = 2$*. The function* $g(z)$ *is shown by the solid line, the majorants by the dashed line. Both majorants are very close to* $g(z)$ *near zero and* μ*.*

It remains the case with x in $[-1,1]$. The approximation on this subset is given by the parabola $-\log(2)\ x^2$ for all three majorants. We note that the parabola is constructed such that it goes through the points $(-1, f(-1) = -\log(2))$, $(0, f(0) = 0)$, $(1, f(1) = -\log(2))$. In addition, we can prove the inequality on $[-1,1]$ by proving that the function

$$h(x) := -\log(2)x^2 - f(x) = -\log(2)x^2 + \log\left(1+x^2\right)$$

is positive on the subset $[0,1]$. (Then $h(x)$ is positive on the subset $[-1,1]$ thanks to the symmetry about zero.) To this aim, we note that $h(0) = h(1) = 0$. Moreover, the first derivative is given by

$$h'(x) = -2\ \log(2)\ x + \frac{2\ x}{1+x^2} = -2\ x\left(\log(2) - \frac{1}{1+x^2}\right).$$

Thus, the roots of $h'(x)$ are at $x_{min} = 0$ and $x_{max} = \sqrt{\frac{1}{\log(2)} - 1}$ (we do not consider the corresponding negative root). As pointed out by the indices, the first root corresponds to a minimum of the function $h(x)$ and the second one to a maximum. Then, $h'(x)$ is positive on $[0, x_{max}]$ and negative on $[x_{max}, 1]$ and the previous assertion about $h(x)$ follows by applying twice the mean value theorem. In fact, we have for $x \in [0, x_{max}]$:

$$h(x) - h(0) = h'(\xi)\ (x-0) \qquad \text{with} \qquad \xi \in (0,x) \subset [0, x_{max}].$$

Thus, $h(x) = h'(\xi)\ (x-0) \geq 0$. On the other hand, we have for $x \in [x_{max}, 1]$:

$$h(1) - h(x) = h'(\xi)\ (1-x) \qquad \text{with} \qquad \xi \in (x,1) \subset [x_{max}, 1].$$

Thus, $h(x) = -h'(\xi)\ (1-x) \geq 0$.

The continuity feature follows easily from the construction of the majorants. Note that the majorants have the same value in the points d_k both coming from the left and from the right side of them. This can be seen also in Figure 2.3. □

Remark 2.9 *We saw in the previous proof that the choice of the points $x = \{-1,1\}$ (respectively $z = \{-c/\sigma_V, c/\sigma_V\}$) to define the subsets B_k was not accidental. These points delimit the concave region of $f(x)$. Also the choice of the points $x = \{-\sqrt{3}, \sqrt{3}\}$ (respectively*

$z = \{-\sqrt{3}\ c/\sigma_V, \sqrt{3}\ c/\sigma_V\}$) has a reason. In fact, we have

$$\begin{aligned}\frac{d}{dx} f''(x) = f'''(x) &= 2\ \frac{2\ x\ (1+x^2)^2 - (x^2-1)\ 2\ (1+x^2)\ 2\ x}{(1+x^2)^4} \\ &= 2\ \frac{(1+x^2)}{(1+x^2)^4}\ \left[2\ x\ (1+x^2) - 4\ x\ (x^2-1)\right] \\ &= -4\ \frac{(1+x^2)}{(1+x^2)^4}\ \left[x\ (x-\sqrt{3})\ (x+\sqrt{3})\right].\end{aligned}$$

The roots of $f'''(x)$ are at 0 and $\pm\sqrt{3}$. It follows that the function $f''(x)$ has maxima at $x = \{-\sqrt{3}, \sqrt{3}\}$ and a minimum at $x = 0$. Since $f''(x)$ describes the curvature of the function $f(x)$, we suggest to take the points where the curvature is maximal to define the subsets B_k.

2.3.5 Summary of the particle filtering recursion with $\widetilde{y}_t$ available

Finally, we put together all results of the previous subsections and we write the algorithm to implement the filtering recursion at time t with $\widetilde{y}_t$ available. As stressed in Remark 2.8, the default majorant can be computed at the beginning of the filtering recursion since it depends only on the value c/σ_V. In addition, some components of the lower and the upper majorants can be computed as well at the beginning since they depend only on c/σ_V.
Thus, the filtering recursion at time t with $\widetilde{y}_t$ available is organized as follows.

Algorithm 2.1 *Particle filtering recursion at time t with $\widetilde{y}_t$ available.*
Assumptions:

- *The fully defined default majorant and the partially defined lower and upper majorants have already been computed and δ is known, see Definition 2.3.*
- *The sample $(z^{(i)}_{(t-p):(t-1)})$ of $Z_{(t-p):(t-1)}|\widetilde{Y}_{1:(t-1)} = \widetilde{y}_{1:(t-1)}$ is known from the filtering recursion at time $t-1$.*

Preliminaries:

1. *Compute the μ_i's, $i = 1, \ldots, N$, as described in Lemma 2.3.*

2. *Choose the majorant for this step (time t).*

   To this end, define
$$\mu = med_i\,(\mu_i)$$
   and select
   - *the default majorant if* $-\sqrt{3}\,\frac{c}{\sigma_V} - \delta \le \mu \le \sqrt{3}\,\frac{c}{\sigma_V} + \delta$,
   - *the lower majorant if* $-\sqrt{3}\,\frac{c}{\sigma_V} - \delta > \mu$,
   - *the upper majorant if* $\mu > \sqrt{3}\,\frac{c}{\sigma_V} + \delta$.

   If the lower or the upper majorant is chosen, compute the components which depend on μ and δ, see Definition 2.3.
3. *Compute the setup to apply the rejection sampling method with the auxiliary index, i.e. find the variables to get the efficient proposal densities $\rho\,(i,z)$ and the optimal distribution $(\tau(i))$.*

   To this end, compute the variables $\overline{\sigma}_k$, $\overline{\mu}_{k,i}$, $\Phi\left(\frac{d_k-\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)$, $\Phi\left(\frac{d_{k-1}-\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)$, $RM_{k,i} := R_{k,i}\cdot M_{k,i}$ for all $k = 1,\ldots,K$ and all $i = 1,\ldots,N$ as described in Lemma 2.3. Then, compute the partial sums of $RM_{k,i}$ over k for all i. In addition, use Lemma 2.4 to find the optimal distribution $(\tau(i))$ and its partial sums.

Begin the construction of the sample $(z^{(l)}_{(t-p+1):t})$, $l = 1,\ldots,N$, of $Z_{(t-p+1):t}|\widetilde{Y}_{1:t} = \widetilde{y}_{1:t}$. Set $l = 1$.

4. *Sample a pair $\left(i^{(l)}, z^{(l)}\right)$ according to the distribution $\tau(i)\rho\,(i,z)$.*

   First, generate the auxiliary index $I^{(l)}$ according to the distribution $(\tau(i))$ and then the variable $Z^{(l)}$ according to the density $\rho\left(i^{(l)}, z\right)$ with $I^{(l)} = i^{(l)}$. The two samplings are carried out by the inversion method. The partial sums of $(\tau(i))$ and the partial sums of $RM_{k,i}$ over k for all i are needed, see also Remark 2.6.
5. *Check the acceptance of the proposed pair $\left(i^{(l)}, z^{(l)}\right)$.*

   For this purpose, generate U uniform on $[0,1]$ and compute the acceptance probability $\pi\left(i^{(l)}, z^{(l)}\right)$ according to Lemma 2.5.
   - *If $U \le \pi\left(i^{(l)}, z^{(l)}\right)$, then accept the pair $\left(i^{(l)}, z^{(l)}\right)$. Return the particle $z^{(l)}_{(t-p+1):t}$ defined by*
$$\begin{aligned} z^{(l)}_{t-p+j} &= z^{(i^{(l)})}_{t-p+j} \qquad \text{for} \quad j = 1,\ldots,p-1,\\ z^{(l)}_t &= \widetilde{y}_t + \sigma_V\; z^{(l)}. \end{aligned}$$

Set $l = l + 1$.
If $l \leq N$, return to step 4. Otherwise stop: all particles have been computed.

- *Else, the pair $\left(i^{(l)}, z^{(l)}\right)$ is not accepted. Return to step 4.*

2.4 Particle filtering recursion with missing $\widetilde{y}_t$

In this section, we go back to the second case that we have distinguished in the implementation of the filtering recursion at time t, that is the case where $\widetilde{y}_t$ is missing. The implementation becomes simpler, as already mentioned in Section 2.3.

If $\widetilde{y}_t$ is missing, we set the filtering density $f_{t|t}\left(z_{(t-p+1):t}|\widetilde{y}_{1:t}\right)$ equal to the prediction density $f_{t|t-1}\left(z_{(t-p+1):t}|\widetilde{y}_{1:(t-1)}\right)$ (we do not have an update step). The integral in the latter density is approximated by particles as described in Section 2.2. Thus, we have

$$f_{t|t}\left(z_{(t-p+1):t}|\widetilde{y}_{1:t}\right) \approx \frac{1}{N}\sum_{i=1}^{N} p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)\Delta\left(z^{(i)}_{(t-p+1):(t-1)}\right)$$

with $(z^{(i)}_{(t-p):(t-1)})$ a sample of $Z_{(t-p):(t-1)}|\widetilde{Y}_{1:(t-1)} = \widetilde{y}_{1:(t-1)}$. Note that we have not dropped $\widetilde{y}_t$ from the previous filtering density notation, although it is missing. We have set the "value" of $\widetilde{y}_t$ to NA (*not available*). In this way, formulae have a better readability.
Similar to the case with $\widetilde{y}_t$ available, the crucial point is to generate a sample $(z^{(l)}_t)$ from the density

$$\widehat{f}_{t|t}\left(z_t|\widetilde{y}_{1:t}\right) = \frac{1}{N}\sum_{i=1}^{N} p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right).$$

Since this density is a mixture of N densities and we have to sample N particles $z^{(l)}_t$ from it, we suggest to use a deterministic mixing for the sampling. I.e. we sample once from each density $p\left(z_t|z^{(i)}_{(t-p):(t-1)}\right)$. Therefore, the filtering recursion at time t with missing $\widetilde{y}_t$ is organized as follows.

Algorithm 2.2 *Particle filtering recursion at time t with missing $\widetilde{y}_t$.*
Assume that the sample $(z^{(i)}_{(t-p):(t-1)})$ of $Z_{(t-p):(t-1)}|\widetilde{Y}_{1:(t-1)} = \widetilde{y}_{1:(t-1)}$ is known from the filtering recursion at time $t-1$.
Then:
For i from 1 to N do:

a) Sample $Z^{(i)}_t$ according to the state density

$$p\left(z_t | z^{(i)}_{(t-p):(t-1)}\right) = \phi\left(\sum_{l=1}^{p} \varphi_l z^{(i)}_{t-l}, \sigma_V\right)(z_t)\,.$$

b) Return the particle $z^{(i)}_{(t-p+1):t}$. The components $z^{(i)}_{t-p+1}, \ldots, z^{(i)}_{t-1}$ are taken from the input filter particle $z^{(i)}_{(t-p):(t-1)}$.

Chapter 3

Smoothing recursion

We have seen in Chapter 2 how we can find the particle filtering recursion for the considered model (2.6) and (2.7). The filtering recursion has a pleasant characteristic: if a new observation $\widetilde{y}_{t'}$ becomes available, we can easily carry on the filtering recursion for this new value. In fact, we only need the previous sample $(z^{(l)}_{(t'-p):(t'-1)})$ besides the observation $\widetilde{y}_{t'}$. This permits to apply the filtering recursion in on-line studies where the observations become available one after the other.
The estimation of the state variables (Z_t) would take advantage if we also knew the future observations $\widetilde{y}_{t'+1}, \widetilde{y}_{t'+2}, \ldots$ in addition to the past observations $\widetilde{y}_{1:t'}$. This is easy to understand if the observed value $\widetilde{y}_{t'}$ is an outlier. In fact, also the sample $(z^{(l)}_{(t'-p+1):t'})$ generated with our filtering method may be influenced by the outlier $\widetilde{y}_{t'}$ although in a less pronounced manner than with the Kalman filtering method. If we knew the future observed values $\widetilde{y}_{t'+1}, \widetilde{y}_{t'+2}, \ldots$, then the presence of the outlier $\widetilde{y}_{t'}$ could be noticed (we recall at this point that the observation errors in (2.7) are assumed to be independent). Thus, we could reduce the influence of the outlier and get a better estimate of the state variable $Z_{t'}$. Of course, such a method can be carried out only off-line, i.e. once all observed values are collected.
This idea leads to *smoothing methods.* We discuss it in this chapter. The considered model is the same as in Chapter 2, see (2.6) and (2.7), and again the parameters in the model are assumed to be known. Now, the whole set of observations $\widetilde{y}_{1:T}$ is available and the goal will be to generate a sample from the density $p\,(z_{1:T}|\widetilde{y}_{1:T})$. To succeed in this, we

will use the filter samples at all times t in $\{1, \dots, T\}$. Thus, the filtering recursion has to be executed first, and all generated particles should be saved. This is a disadvantage since it takes some time and memory.
On the other hand, the smoothing algorithm will improve the estimates of the state variables (Z_t) and, above all, it permits to derive fast and reliable maximum likelihood estimating methods for the parameters if they are no more assumed to be given (see Chapter 4). But a prerequisite for the maximum likelihood methods is that the sampling from $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$ is fast to execute. For this reason, efficiency of the smoothing algorithm is important.

The chapter is organized as follows. The beginning is dedicated to find the exact density $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$. Then two methods are presented to sample from this density. The first one is an adjustment of the implemented particle filtering method. Unfortunately, the resulting algorithm can be very slow since its complexity is of order TN^2 (T the length of the observed series (Y_t) and N the size of the generated sample). The second method is an improvement of the first one to speed up the algorithm. In fact, it has a complexity of order $TN\log(N)$. This can be achieved at the expense of a more difficult implementation.

3.1 Exact density $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$

In this section, we compute the density $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$ for the considered model (2.6) and (2.7). The aim is to write it in a form which permits an easy sampling from it.
First, it is helpful to illustrate with a graph the dependence structure of the state variable Z_t when all observations $\widetilde{Y}_{1:T}$ are known, see Figure 3.1. The used conditional independence properties can be easily read from it.

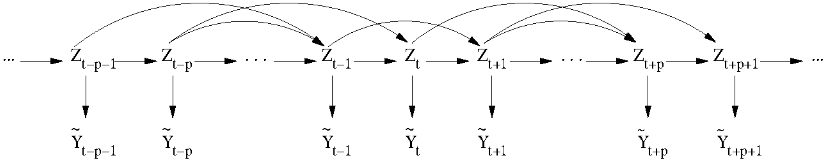

Figure 3.1: *Dependence structure of the state variable Z_t when all observations $\widetilde{Y}_{1:T}$ are known.*

The density $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$ can be written as

$$\begin{aligned}
p\left(z_{1:T}|\widetilde{y}_{1:T}\right) &= p\left(z_{(T-p+1):T}|\widetilde{y}_{1:T}\right) p\left(z_{1:(T-p)}|z_{(T-p+1):T}, \widetilde{y}_{1:T}\right) \\
&= p\left(z_{(T-p+1):T}|\widetilde{y}_{1:T}\right) \prod_{t=1}^{T-p} p\left(z_t|z_{(t+1):T}, \widetilde{y}_{1:T}\right) \\
&= p\left(z_{(T-p+1):T}|\widetilde{y}_{1:T}\right) \prod_{t=1}^{T-p} p\left(z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right).
\end{aligned}$$

The last equality follows since Z_t is independent of both $\widetilde{Y}_{(t+1):T}$ and $Z_{(t+p+1):T}$ if $Z_{(t+1):(t+p)}$ are given.

Up to now, we have used a classic approach to compute the density $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$. In fact, we use the filtering density $p\left(z_{(T-p+1):T}|\widetilde{y}_{1:T}\right)$ and we go backward in time with transition densities $p\left(z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$. The next step is to compute these transition densities. To this aim, we use twice Bayes' theorem, we introduce the past state variables $Z_{(t-p):(t-1)}$ and we argue with similar conditional independence properties as before. Explicitly,

$$\begin{aligned}
p\left(z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right) &= \frac{p\left(\widetilde{y}_t, z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right)}{p\left(\widetilde{y}_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right)} \\
&= \frac{p\left(\widetilde{y}_t|z_t, z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right) p\left(z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right)}{p\left(\widetilde{y}_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right)} \qquad (3.1) \\
&= \frac{p\left(\widetilde{y}_t|z_t\right) p\left(z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right)}{p\left(\widetilde{y}_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right)} \\
&= \frac{p\left(\widetilde{y}_t|z_t\right) \int \cdots \int p\left(z_{(t-p):(t-1)}, z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right) dz_{t-p} \ldots dz_{t-1}}{p\left(\widetilde{y}_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right)} \\
&= \frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right)} \cdot \\
&\cdot \int \cdots \int p\left(z_t|z_{(t-p):(t-1)}, z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right) \cdot \\
&\qquad \cdot p\left(z_{(t-p):(t-1)}|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right) dz_{t-p} \ldots dz_{t-1}
\end{aligned}$$

$$= \frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:(t-1)}\right) p\left(z_{(t+1):(t+p)}|\widetilde{y}_{1:(t-1)}\right)} \cdot$$

$$\cdot \int \cdots \int p\left(z_t|z_{(t-p):(t-1)}, z_{(t+1):(t+p)}\right) p\left(z_{(t-p):(t-1)}, z_{(t+1):(t+p)}|\widetilde{y}_{1:(t-1)}\right)$$

$$dz_{t-p} \ldots dz_{t-1}$$

$$= \frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t, z_{(t+1):(t+p)}|\widetilde{y}_{1:(t-1)}\right)} \cdot$$

$$\cdot \int \cdots \int p\left(z_t|z_{(t-p):(t-1)}, z_{(t+1):(t+p)}\right) p\left(z_{(t+1):(t+p)}|z_{(t-p):(t-1)}, \widetilde{y}_{1:(t-1)}\right) \cdot$$

$$\cdot\, p\left(z_{(t-p):(t-1)}|\widetilde{y}_{1:(t-1)}\right) dz_{t-p} \ldots dz_{t-1}$$

$$= \frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t, z_{(t+1):(t+p)}|\widetilde{y}_{1:(t-1)}\right)} \cdot$$

$$\cdot \int \cdots \int p\left(z_t|z_{(t-p):(t-1)}, z_{(t+1):(t+p)}\right) p\left(z_{(t+1):(t+p)}|z_{(t-p):(t-1)}\right) \cdot$$

$$\cdot\, p\left(z_{(t-p):(t-1)}|\widetilde{y}_{1:(t-1)}\right) dz_{t-p} \ldots dz_{t-1}.$$

Two comments about the last expression. The integral can be seen as expected value with respect to the variables $Z_{(t-p):(t-1)}|\widetilde{Y}_{1:(t-1)}$. We will use this feature to approximate this integral, see Section 3.2. Moreover, the denominator does not depend on Z_t. It is only the normalizing constant.

We consider closer the product

$$p\left(z_t|z_{(t-p):(t-1)}, z_{(t+1):(t+p)}\right) p\left(z_{(t+1):(t+p)}|z_{(t-p):(t-1)}\right)$$

in the last expression. This will be useful for Section 3.2.

Lemma 3.1 *Let (Z_t) be a stationary Gaussian AR process as described by the state equation (2.6).*
Then:

$$p\left(z_t|z_{(t-p):(t-1)}, z_{(t+1):(t+p)}\right)\ p\left(z_{(t+1):(t+p)}|z_{(t-p):(t-1)}\right) =$$
$$= w^{(-,+)}\ \phi\left(\mu^{(-)} + \mu^{(+)}, \widetilde{\sigma}\right)(z_t)$$

with

$$\varphi := \left(\varphi_1, \ldots, \varphi_p\right)',$$

$$\begin{aligned}
\widetilde{\sigma}^2 &:= \frac{\sigma_V^2}{1+\varphi'\varphi},\\
\mu^{(-)} &:= \mu^{(-)}(z_{(t-p):(t-1)}) = \left(1+\varphi'\varphi\right)^{-1}\sum_{l=1}^{p}\left(\varphi_l - \sum_{s=1}^{p-l}\varphi_s\varphi_{s+l}\right)z_{t-l},\\
\mu^{(+)} &:= \mu^{(+)}(z_{(t+1):(t+p)}) = \left(1+\varphi'\varphi\right)^{-1}\sum_{l=1}^{p}\left(\varphi_l - \sum_{s=1}^{p-l}\varphi_s\varphi_{s+l}\right)z_{t+l},\\
\Sigma &:= \sigma_V^2\left(\mathbb{I}+\varphi\varphi'\right),\\
w^{(-,+)} &:= w^{(-,+)}\left(z_{(t-p):(t-1)}, z_{(t+1):(t+p)}\right)\\
&= \phi\left(v^{(-)},\Sigma^{0.5}\right)\left(v^{(+)}\right)\\
&= \frac{1}{(2\pi)^{p/2}\sigma_V^p\sqrt{1+\varphi'\varphi}}\exp\left[-\frac{1}{2}\left(v^{(+)}-v^{(-)}\right)'\Sigma^{-1}\left(v^{(+)}-v^{(-)}\right)\right].
\end{aligned}$$

The components of the vectors

$$v^{(-)} := \left(v^{(-)(1)},\ldots,v^{(-)(p)}\right)' \quad \text{and} \quad v^{(+)} := \left(v^{(+)(1)},\ldots,v^{(+)(p)}\right)'$$

are given by

$$\begin{aligned}
v^{(-)(l)} &:= v^{(-)(l)}(z_{(t-p):(t-1)}) = \sum_{s=l+1}^{p}\varphi_s z_{t+l-s} + \varphi_l\sum_{s=1}^{p}\varphi_s z_{t-s},\\
v^{(+)(l)} &:= v^{(+)(l)}(z_{(t+1):(t+p)}) = z_{t+l} - \sum_{s=1}^{l-1}\varphi_s z_{t+l-s}.
\end{aligned}$$

The inverse of the matrix Σ can be computed explicitly. We have

$$\Sigma^{-1} = \frac{1}{\sigma_V^2}\left(\mathbb{I} - \frac{1}{1+\varphi'\varphi}\,\varphi\varphi'\right).$$

Remark 3.1 *Since the matrix Σ^{-1} is symmetric positive definite, we can decompose it with the Cholesky method. I.e. there exists an upper triangular matrix R such that*

$$\Sigma^{-1} = R'R.$$

Then, $w^{(-,+)}$ *can be rewritten as*

$$\begin{aligned} w^{(-,+)} &= k \;\exp\left[-\frac{1}{2}\left(v^{(+)} - v^{(-)}\right)' \;\Sigma^{-1}\; \left(v^{(+)} - v^{(-)}\right)\right] \\ &= k \;\exp\left[-\frac{1}{2}\left(v^{(+)} - v^{(-)}\right)' \;R'R\; \left(v^{(+)} - v^{(-)}\right)\right] \\ &= k \;\exp\left[-\frac{1}{2}\left(R\left(v^{(+)} - v^{(-)}\right)\right)' \;R\left(v^{(+)} - v^{(-)}\right)\right] \\ &= \phi\left(Rv^{(-)}, \mathbb{I}\right)\left(Rv^{(+)}\right) \end{aligned}$$

where k *denotes the normalizing constant in the definition of* $w^{(-,+)}$*. This is a very fine interpretation since the distances measured with an identity covariance matrix are actually Euclidean distances. We will use this feature to implement the second smoothing algorithm, see Section 3.4.*

Proof: The proof follows by showing the two assertions:

1. $p\left(z_t | z_{(t-p):(t-1)}, z_{(t+1):(t+p)}\right) = \phi\left(\mu^{(-)} + \mu^{(+)}, \widetilde{\sigma}\right)(z_t)$,
2. $p\left(v^{(+)} | z_{(t-p):(t-1)}\right) = \phi\left(v^{(-)}, \Sigma^{0.5}\right)\left(v^{(+)}\right)$ and
   $p\left(z_{(t+1):(t+p)} | z_{(t-p):(t-1)}\right) = p\left(v^{(+)} | z_{(t-p):(t-1)}\right)$.

Let us begin with the first assertion. The spectral density of (Z_t) is given by

$$\begin{aligned} f(\nu) &= \frac{\sigma_V^2}{\left|1 - \sum\limits_{l=1}^{p} \varphi_l \exp\left(-i2\pi\nu l\right)\right|^2} \\ &= \frac{\sigma_V^2}{\left(1 - \sum\limits_{l=1}^{p} \varphi_l \exp\left(-i2\pi\nu l\right)\right)\overline{\left(1 - \sum\limits_{l=1}^{p} \varphi_l \exp\left(-i2\pi\nu l\right)\right)}} \\ &= \frac{\sigma_V^2}{1 - 2\sum\limits_{l=1}^{p} \varphi_l \cos(2\pi\nu l) + \sum\limits_{l=1}^{p} \varphi_l^2 + \sum\limits_{l=1}^{p} \sum\limits_{\substack{s=1 \\ s \neq l}}^{p} \varphi_l \varphi_s \exp\left(i2\pi\nu(s-l)\right)} \\ &= \frac{\sigma_V^2}{1 + \sum\limits_{l=1}^{p} \varphi_l^2 - 2\sum\limits_{l=1}^{p} \varphi_l \cos(2\pi\nu l) + \sum\limits_{l=1}^{p} \sum\limits_{s=1}^{p-l} \varphi_s \varphi_{s+l} \left(\exp\left(i2\pi\nu l\right) + \exp\left(-i2\pi\nu l\right)\right)} \end{aligned}$$

$$= \frac{\sigma_V^2}{1+\sum\limits_{l=1}^{p}\varphi_l^2 - 2\sum\limits_{l=1}^{p}\cos(2\pi\nu l)\left(\varphi_l - \sum\limits_{s=1}^{p-l}\varphi_s\varphi_{s+l}\right)}.$$

Thus, the first assertion follows using standard results for Gaussian Markov fields.

Now, the second assertion. We use first the definition of the AR(p) process (Z_t) to write

$$\begin{aligned} Z_{t+l} &= \sum_{s=1}^{p}\varphi_s Z_{t+l-s} + V_{t+l} \\ &= \sum_{s=1}^{l-1}\varphi_s Z_{t+l-s} + \varphi_l\left(\sum_{s=1}^{p}\varphi_s Z_{t-s} + V_t\right) + \sum_{s=l+1}^{p}\varphi_s Z_{t+l-s} + V_{t+l}. \end{aligned}$$

Thus

$$Z_{t+l} - \sum_{s=1}^{l-1}\varphi_s Z_{t+l-s} = \varphi_l\sum_{s=1}^{p}\varphi_s Z_{t-s} + \sum_{s=l+1}^{p}\varphi_s Z_{t+l-s} + \varphi_l V_t + V_{t+l},$$

i.e.

$$v^{(+)(l)} = v^{(-)(l)} + \varphi_l V_t + V_{t+l}.$$

Since $v^{(-)(l)}$ is a linear combination of $Z_{(t-p):(t-1)}$ and both errors V_t and V_{t+l} do not depend on these past state variables, we find considering all l that

$$v^{(+)}|Z_{(t-p):(t-1)} \sim \mathcal{N}(v^{(-)}, \Sigma)$$

with covariance matrix $\Sigma = \sigma_V^2\left(\mathbb{I} + \varphi\varphi^{'}\right)$. The inverse of Σ is given by

$$\Sigma^{-1} = \frac{1}{\sigma_V^2}\left(\mathbb{I} - \frac{1}{1+\varphi^{'}\varphi}\,\varphi\varphi^{'}\right).$$

In fact,

$$\begin{aligned} \Sigma\Sigma^{-1} &= \sigma_V^2\left(\mathbb{I} + \varphi\varphi^{'}\right)\frac{1}{\sigma_V^2}\left(\mathbb{I} - \frac{1}{1+\varphi^{'}\varphi}\,\varphi\varphi^{'}\right) \\ &= \mathbb{I} + \left(1 - \frac{1}{1+\varphi^{'}\varphi} - \frac{\varphi^{'}\varphi}{1+\varphi^{'}\varphi}\right)\varphi\varphi^{'} = \mathbb{I} \end{aligned}$$

where we use that

$$\varphi\varphi^{'}\varphi\varphi^{'} = \varphi\left(\varphi^{'}\varphi\right)\varphi^{'} = \left(\varphi^{'}\varphi\right)\varphi\varphi^{'}$$

since $\varphi^{'}\varphi$ is a scalar.
Moreover, we can compute explicitly the normalizing constant of the above multidimensional normal distribution. In general, it is given by $(2\pi)^{-p/2}\left(det(\Sigma)\right)^{-1/2}$. But, since Σ is positive definite, it can be diagonalized and $det(\Sigma)$ is the product of the eigenvalues. Recalling that the eigenvalues λ are defined such that $(\Sigma - \lambda\mathbb{I})\,v = 0$ for vectors v different from the zero vector, we see that the eigenvalues are given by σ_V^2 with multiplicity $p-1$ and by $\sigma_V^2\left(1+\varphi^{'}\varphi\right)$. In fact, with $\lambda = \sigma_V^2$, the previous defining equation becomes

$$0 = (\Sigma - \lambda\mathbb{I})\,v = \sigma_V^2\left[\left(\mathbb{I}+\varphi\varphi^{'}\right)-\mathbb{I}\right]v = \sigma_V^2\varphi\varphi^{'}v$$

which can be fulfilled by $p-1$ linear independent vectors v since the rank of $\varphi\varphi^{'}$ is 1 (take v orthogonal). In addition, with $\lambda = \sigma_V^2\left(1+\varphi^{'}\varphi\right)$ and $v = \varphi$, we find

$$\begin{aligned}(\Sigma - \lambda\mathbb{I})\,v &= \sigma_V^2\left[\mathbb{I}+\varphi\varphi^{'}-\left(1+\varphi^{'}\varphi\right)\mathbb{I}\right]\varphi \\ &= \sigma_V^2\left[\varphi+\varphi\left(\varphi^{'}\varphi\right)-\varphi-\left(\varphi^{'}\varphi\right)\varphi\right] = 0.\end{aligned}$$

Therefore, the normalising constant can be written as

$$\begin{aligned}(2\pi)^{-p/2}\left(det(\Sigma)\right)^{-1/2} &= (2\pi)^{-p/2}\left[\sigma_V^{2(p-1)}\,\sigma_V^2\left(1+\varphi^{'}\varphi\right)\right]^{-1/2} \\ &= (2\pi)^{-p/2}\,\sigma_V^{-p}\left(1+\varphi^{'}\varphi\right)^{-1/2}.\end{aligned}$$

The assertion that $p\left(v^{(+)}|z_{(t-p):(t-1)}\right) = \phi\left(v^{(-)},\Sigma^{0.5}\right)\left(v^{(+)}\right)$ and also some other features are proved.
Finally, the relationship between $Z_{(t+1):(t+p)}$ and $v^{(+)}$ is linear with functional determinant equal 1. Thus, it follows that

$$p\left(z_{(t+1):(t+p)}|z_{(t-p):(t-1)}\right) = p\left(v^{(+)}|z_{(t-p):(t-1)}\right).$$

□

We recapitulate briefly the results of this section. The density $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$ can be written as

$$p\left(z_{1:T}|\widetilde{y}_{1:T}\right) = p\left(z_{(T-p+1):T}|\widetilde{y}_{1:T}\right) \prod_{t=1}^{T-p} p\left(z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right). \tag{3.2}$$

In addition,

$$\begin{aligned} p\left(z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right) = &\frac{p_{VII}\left(m, c, \widetilde{y}_t\right)\left(z_t\right)}{p\left(\widetilde{y}_t, z_{(t+1):(t+p)}|\widetilde{y}_{1:(t-1)}\right)} \cdot \\ &\cdot \int \cdots \int w^{(-,+)} \cdot \phi\left(\mu^{(-)} + \mu^{(+)}, \widetilde{\sigma}\right)\left(z_t\right) \cdot p\left(z_{(t-p):(t-1)}|\widetilde{y}_{1:(t-1)}\right) \\ &\qquad dz_{t-p} \ldots dz_{t-1} \end{aligned} \tag{3.3}$$

with $\mu^{(-)}$, $\mu^{(+)}$, $w^{(-,+)}$ and $\widetilde{\sigma}$ as in Lemma 3.1.

3.2 Particle smoothing method

The density (3.3) cannot be computed in closed form for the considered model (2.6) and (2.7) and, consequently, the density $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$ is also not available in closed form. What we can do is to approximate $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$ by a sample $(z^{(j)}_{1:T})$ from it. The results (3.2) and (3.3) suggest the following strategy. We start with a sample $(z^{(j)}_{(T-p+1):T})$ from $p\left(z_{(T-p+1):T}|\widetilde{y}_{1:T}\right)$. Then, we go backward sampling the other components from the approximate transition densities $\widehat{p}\left(z_t|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$. As a starting sample $(z^{(j)}_{(T-p+1):T})$ we use the last sample computed in the particle filtering recursion. On the other hand, the transition densities $\widehat{p}\left(z_t|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$ are found approximating (3.3) by the Monte Carlo method. We find

$$\widehat{p}\left(z_t|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right) \propto p_{VII}\left(m, c, \widetilde{y}_t\right)\left(z_t\right) \cdot \sum_{i=1}^{N} w_{i,j} \cdot \phi\left(\mu^{(-)}_i + \mu^{(+)}_j, \widetilde{\sigma}\right)\left(z_t\right) \tag{3.4}$$

with $(z^{(i)}_{(t-p):(t-1)})$ a sample of $Z_{(t-p):(t-1)}|\widetilde{Y}_{1:(t-1)} = \widetilde{y}_{1:(t-1)}$. In addition,

$$\mu_i^{(-)} := \mu^{(-)}\left(z^{(i)}_{(t-p):(t-1)}\right), \tag{3.5}$$

$$\mu_j^{(+)} := \mu^{(+)}\left(z^{(j)}_{(t+1):(t+p)}\right), \tag{3.6}$$

$$v_i^{(-)} := v^{(-)}\left(z^{(i)}_{(t-p):(t-1)}\right), \tag{3.7}$$

$$v_j^{(+)} := v^{(+)}\left(z^{(j)}_{(t+1):(t+p)}\right), \tag{3.8}$$

$$\begin{aligned} w_{i,j} &:= w^{(-,+)}\left(z^{(i)}_{(t-p):(t-1)}, z^{(j)}_{(t+1):(t+p)}\right) \\ &= k\ \exp\left(-\frac{1}{2}\left\|Rv_j^{(+)} - Rv_i^{(-)}\right\|_2^2\right) \end{aligned} \tag{3.9}$$

where $\mu^{(-)}$, $\mu^{(+)}$, $v^{(-)}$, $v^{(+)}$ and $w^{(-,+)}$ are defined as in Lemma 3.1, R as in Remark 3.1.

As we can see, we need all generated filter samples to implement this idea and not only the sample at time T. Thus, the filtering recursion should be executed first and all generated samples should be saved. In addition, we note that we sample from the approximate transition density (3.4) only once for each particle $z^{(j)}_{(t+1):(t+p)}$. In fact, the aim is to produce a sample $(z^{(j)}_{1:T})$ of $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$. The density (3.4) is similar to the target density in the particle filtering steps, compare (3.4) with (2.13). Then, the two sampling methods discussed in connection with the filtering recursion can be applied in principle also here, see their definition in Subsections 2.2.1 and 2.2.2. However, the sampling importance resampling method is not the most appropriate technique in this case since we need only one value from each density $\widehat{p}\left(z_t|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$. For the same reason, we have to be careful in the application of the rejection method with the auxiliary index. The needed proposal densities $\rho_j\left(i, z_t\right)$ and the auxiliary distribution $(\tau_j(i))$ depend on the considered particle $z^{(j)}_{(t+1):(t+p)}$. But it is not convenient to compute these distributions (and the related partial sums to sample from them) for every j since only one particle $z_t^{(j)}$ has to be generated. In fact, the resulting algorithm would be too slow. Thus, we have to find a way to generate more $z_t^{(j)}$'s without changing too much the distribution setup or find an approach where the full setup is computed only if it is needed.

These are the leading ideas of the two presented methods to sample from (3.4). The first approach is similar to the one used to implement the filtering recursion. The idea is to use the same setup to sample several $z_t^{(j)}$'s. Unfortunately, the algorithm can be very slow since it has a complexity of order TN^2. The second method refines the sampling strategy of the first one by grouping the particles $(z_{(t-p):(t-1)}^{(i)})$ according to a specified criterion and by introducing a pretesting for the particle acceptance. Thus, the full distribution setup is computed only if the pretesting is passed. The resulting method will be faster than the first one since it has a complexity of order $TN\log(N)$. But some new theoretical considerations are required to justify it.

3.3 Particle smoothing recursion: method 1

In this section, we explain the first method to sample once from $\widehat{p}\left(z_t|z_{(t+1):(t+p)}^{(j)}, \widetilde{y}_{1:t}\right)$ for each particle $z_{(t+1):(t+p)}^{(j)}$. As noted earlier, the previous density has a similar structure as the target density in the filtering steps: it is proportional to the product of the Pearson type VII observation density with a mixture of weighted normal densities, see (2.13). The differences are in the presence of the weights $w_{i,j}$ in the mixture and in the dependence of the normal densities on both the past filter sample $(z_{(t-p):(t-1)}^{(i)})$ and the future smoothing sample $(z_{(t+1):(t+p)}^{(j)})$. In addition, we stress again that only one particle $z_t^{(j)}$ has to be generated from each $\widehat{p}\left(z_t|z_{(t+1):(t+p)}^{(j)}, \widetilde{y}_{1:t}\right)$. Thus, it is intuitive to apply the same sampling technique as in the filtering recursion and to adapt it to consider the mentioned differences. Again, we have to distinguish two cases: $\widetilde{y}_t$ available or missing. First, we consider the case with known $\widetilde{y}_t$.

3.3.1 Recursion with $\widetilde{y}_t$ available

The main ideas to implement the filtering recursion at time t can be summarized as follows. We applied the rejection method with the auxiliary index to generate the required sample. Moreover, we approximated the logarithm of the problematic heavy-tailed observation density by a clever majorant to succeed in constructing the efficient proposal densities $\rho(i, z)$. We suggested to reformulate the sampling task by translating and rescaling the state variable Z_t, see Lemma 2.2. In this way, we speeded up the algorithm since we avoided to compute the majorant in each filtering step.

We proceed in the same way:

Lemma 3.2 *Consider the model (2.6) and (2.7) and let Z_t be a random variable with probability density function $\widehat{p}\left(z_t|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$ as in (3.4).*
Then the random variable Z defined by

$$Z = \frac{Z_t - \widetilde{y}_t}{\widetilde{\sigma}}$$

has density

$$\widehat{p}\left(z|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right) \propto p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \cdot \sum_{i=1}^{N} w_{i,j} \cdot \phi\left(\frac{\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z) \tag{3.10}$$

with $\widetilde{\sigma}$, $\mu_i^{(-)}$, $\mu_j^{(+)}$ and $w_{i,j}$ as in (3.4). In addition, the normalizing constant depends only on $z^{(j)}_{(t+1):(t+p)}$.

Proof: The proof is similar to the previous one, see Lemma 2.2. Substitute the expected values and the standard deviation in the normal densities by the new ones ($\mu_i^{(-)} + \mu_j^{(+)}$ and $\widetilde{\sigma}$, respectively). Note that the additional weights $w_{i,j}$ do not affect the proof since they are independent of Z_t. □

The Pearson type VII density in Lemma 3.2 is no more dependent on $\widetilde{Y}_t$. We proceed for each $j = 1, \ldots, N$ as follows:

- sample one $z^{(j)}$ from the density $\widehat{p}\left(z|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$ defined as in Lemma 3.2,
- set $z_t^{(j)} = \widetilde{y}_t + \widetilde{\sigma}\; z^{(j)}$.

Thus, we can restrict the discussion to the application of the rejection method with the auxiliary index to the sampling of $(z^{(j)})$. First, as in the filtering case, we look for efficient proposal densities $\rho_j(i, z)$ and an optimal distribution $(\tau_j(i))$. Then, we compute the acceptance probabilities of the proposed pairs (i, z) such that, finally, we can write the smoothing recursion for the case with $\widetilde{y}_t$ available. In the discussion, we will stress the modifications needed in comparison to the filtering algorithm.

In Subsection 2.2.1, we saw that each density $\rho_j\left(i,z\right)$ should be a good proposal distribution for the density proportional to

$$p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \cdot w_{i,j}\ \phi\left(\frac{\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)$$

or, equivalently, proportional to

$$p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \cdot \phi\left(\frac{\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)$$

since the weights $w_{i,j}$ are independent of Z. If we would proceed as in the filtering case, we would construct the proposals $\rho_j\left(i,z\right)$ approximating the logarithm of the Pearson type VII density by a majorant. Unfortunately, the densities $\rho_j\left(i,z\right)$ would also depend on the smoothing particles $(z_{(t+1):(t+p)}^{(j)})$ since the latter are used to define the $\mu_j^{(+)}$'s. Thus, we should compute all densities for each j, although we sample only one $Z^{(j)}$. Clearly, the resulting algorithm would be very slow, since the construction of the distribution setup requires many calculations as seen in Algorithm 2.1.

The key observation is the following: if some $\mu_j^{(+)}$'s are near each other, then the corresponding proposals $\rho_j\left(i,z\right)$ will be very similar. Thus, we propose the following recursive procedure. The first proposals $\rho\left(i,z\right)$ are constructed with $\mu_j^{(+)}$ equal to the minimal value of all $\mu_j^{(+)}$'s. We use these proposals $\rho\left(i,z\right)$ to sample possibly several particles $z^{(j')}$. We begin with the indices j' with values $\mu_{j'}^{(+)}$ nearest the chosen $\mu_j^{(+)}$ and then we move away until the densities $\rho\left(i,z\right)$ become too bad proposals. When this happens, a new setup of proposals $\rho\left(i,z\right)$ is constructed with $\mu_j^{(+)}$ equal to the value $\mu_{j''}^{(+)}$, j'' the index where the setup change takes place. The sampling of $z^{(j')}$ goes on from this index j'' and the procedure is iterated until all particles $(z^{(j')})$ are generated. What we need to apply this idea is a break criterion which indicates when the densities $\rho\left(i,z\right)$ become too bad proposals and they should be updated. An intuitive one is given by the number of rejections to sample $z^{(j')}$. When this number exceeds a chosen maximal value, then we have a hint that the proposals are no more appropriate to sample $z^{(j')}$ and new ones have to be computed. Of course, we expect more rejections with this approach than with the one where the proposals are computed for each j, since the used proposals are now nearly optimal. On the other hand, we compute only few updates of all distributions. Therefore, the resulting algorithm will be faster since the proposal of some more pairs (i, z) is cheaper than the computation of a whole distribution setup.

Explicitly, we denote by $\overline{\mu_j^{(+)}}$ the used $\mu_j^{(+)}$ and we define the majorant equivalently to the filtering case, see Definition 2.2.

Definition 3.1 *For a given integer value K, we choose parameters α_k, β_k, γ_k and d_k such that the majorant of the function*

$$g(z) := -\log\left[1 + \left(\frac{z}{c/\widetilde{\sigma}}\right)^2\right]$$

is defined by

$$g(z) \leq \sum_{k=1}^{K} \left(\alpha_k + \beta_k z + \gamma_k z^2\right) \mathbb{I}_{\{z \in B_k\}} \qquad \forall z \in \mathbb{R}$$

with $B_k := (d_{k-1}, d_k]$, $-\infty =: d_0 < d_1 < \cdots < d_K := \infty$. The parameters α_k, β_k, γ_k and d_k are not allowed to depend on z. log *denotes the natural logarithm.*

The remarks done in the filtering case remain also valid here. Again, the construction of the majorant is postponed and we find first the proposals $\rho\,(i, z)$. As said, they should be good proposal distributions for the densities proportional to

$$p_{VII}\,(m, c/\widetilde{\sigma}, 0)\,(z) \cdot \phi\left(\frac{\mu_i^{(-)} + \overline{\mu_j^{(+)}} - \widetilde{y}_t}{\widetilde{\sigma}}, \sqrt{\lambda}\right)(z)$$

where now $\mu_j^{(+)}$ is substituted by $\overline{\mu_j^{(+)}}$. As a consequence of this approximation, we should also modify the variance of the normal densities to succeed in the construction of the distribution $(\tau_j(i))$ and in the evaluation of the acceptance probabilities. For this reason we have introduced the new variance λ. λ cannot depend on the smoothing sample $(z_{(t+1):(t+p)}^{(j)})$ otherwise we reintroduce its dependence in the proposal densities. In addition, we assume that λ does not depend neither on the filter sample $(z^{(i)}_{(t-p):(t-1)})$ nor on z to avoid additional difficulties in the following computations. For the moment, the optimal choice of λ remains open. Note that λ is one of the first thing to compute in the smoothing steps since it influences the construction of the needed distributions.

Lemma 3.3 *Let the proposal densities $\rho\,(i,z)$ be chosen as*

$$\rho\,(i,z) = \sum_{k=1}^{K} \frac{R_{k,i} M_{k,i}}{\sum_{l=1}^{K} R_{l,i} M_{l,i}} \frac{\exp\left[-\frac{1}{2}\left(\frac{z-\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)^2\right]}{M_{k,i}} \, \mathbb{I}_{\{z \in B_k\}}$$

with

$$\overline{\sigma}_k := \sqrt{\frac{1}{\frac{1}{\lambda} - 2m\gamma_k}},$$

$$\overline{\mu}_{k,i} := \overline{\sigma}_k^2 \left(\frac{1}{\lambda} \frac{\mu_i^{(-)} + \overline{\mu_j^{(+)}} - \widetilde{y}_t}{\widetilde{\sigma}} + m\beta_k \right),$$

$$R_{k,i} := \exp\left[\frac{1}{2}\left(\frac{\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)^2 - \frac{1}{2\lambda}\left(\frac{\mu_i^{(-)} + \overline{\mu_j^{(+)}} - \widetilde{y}_t}{\widetilde{\sigma}}\right)^2 + m\alpha_k \right],$$

$$M_{k,i} := \sqrt{2\pi}\;\overline{\sigma}_k \left[\Phi\left(\frac{d_k - \overline{\mu}_{k,i}}{\overline{\sigma}_k}\right) - \Phi\left(\frac{d_{k-1} - \overline{\mu}_{k,i}}{\overline{\sigma}_k}\right) \right].$$

The parameters α_k, β_k, γ_k and d_k are as in Definition 3.1; $\widetilde{\sigma}$ is as in Lemma 3.1; $\mu_i^{(-)}$ and $\mu_j^{(+)}$ are as in (3.5) and (3.6); m and c are the parameters of the observation error distribution in the considered model (2.6) and (2.7). Finally, $\Phi\,(x)$ is the cumulative $\mathcal{N}(0,1)$ distribution function.
Then

$$p_{VII}\,(m, c/\widetilde{\sigma}, 0)\,(z) \cdot \phi\left(\frac{\mu_i^{(-)} + \overline{\mu_j^{(+)}} - \widetilde{y}_t}{\widetilde{\sigma}}, \sqrt{\lambda}\right)(z) \leq k_1 \cdot \left(\sum_{l=1}^{K} R_{l,i} M_{l,i}\right) \cdot \rho\,(i,z) \tag{3.11}$$

with

$$k_1 := \frac{\Gamma(m)\;\widetilde{\sigma}}{\sqrt{2\lambda}\;\pi\;c\;\Gamma(m-0.5)}.$$

Remark 3.2 *Most remarks of the corresponding lemma in the filtering recursion are also valid here, see Remark 2.6. Especially, the sampling of Z according to $\rho\,(i,z)$ can be carried out as explained there.*

Proof: The proof is the same as in the filtering case, see Lemma 2.3. We substitute σ_V by $\widetilde{\sigma}$, $\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)}$ by $\mu_i^{(-)}+\overline{\mu_j^{(+)}}$ and μ_i by $\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}$. In addition, we use the new definitions of $\overline{\sigma}_k$, $\overline{\mu}_{k,i}$, $R_{k,i}$ and k_1 given here. □

We need an interim result before looking for the optimal distribution $(\tau_j(i))$. We have to investigate the relationship between the true normal densities in (3.10) and the approximate ones used to construct the proposals $\rho\,(i,z)$.

Lemma 3.4 *For $\lambda \neq 1$, we have*

$$\frac{\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}},\sqrt{\lambda}\right)(z)}=\sqrt{\lambda}\;\exp\left[\frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)}-\overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2\right]\cdot$$

$$\cdot\exp\left\{-\frac{1}{2}\Big(1-\frac{1}{\lambda}\Big)\left[z-\Big(1-\frac{1}{\lambda}\Big)^{-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\lambda\widetilde{\sigma}}\right)\right]^2\right\}.$$

In addition, it follows for $\lambda > 1$ that

$$\frac{\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}},\sqrt{\lambda}\right)(z)}\leq\sqrt{\lambda}\;\exp\left[\frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)}-\overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2\right]$$

and the minimum of the function

$$h(\lambda):=\sqrt{\lambda}\;\exp\left[\frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)}-\overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2\right]$$

is taken at $\lambda_j=\left(\frac{\sqrt{e_j}+\sqrt{e_j+4}}{2}\right)^2$ *with* $e_j:=\left(\frac{\mu_j^{(+)}-\overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2$.

Proof: The proof requires only algebraic calculations.
Let us begin with the first assertion. The strategy is to put together the two normal densities, to complete the square in the exponent and to simplify it. I.e.:

$$\frac{\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}},\sqrt{\lambda}\right)(z)}=$$

$$=\sqrt{\lambda}\exp\left[-\frac{1}{2}\left(z-\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2+\frac{1}{2\lambda}\left(z-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2\right].$$

It follows by completing the square in the exponent:

$$-\frac{1}{2}\left(z-\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2+\frac{1}{2\lambda}\left(z-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2$$

$$=-\frac{1}{2}\left(1-\frac{1}{\lambda}\right)\left[z^2-2z\left(1-\frac{1}{\lambda}\right)^{-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\lambda\widetilde{\sigma}}\right)\right]-$$

$$-\frac{1}{2}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2+\frac{1}{2\lambda}\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2$$

$$=-\frac{1}{2}\left(1-\frac{1}{\lambda}\right)\left[z-\left(1-\frac{1}{\lambda}\right)^{-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\lambda\widetilde{\sigma}}\right)\right]^2+$$

$$+\frac{1}{2}\left(1-\frac{1}{\lambda}\right)^{-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\lambda\widetilde{\sigma}}\right)^2-$$

$$-\frac{1}{2}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2+\frac{1}{2\lambda}\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2.$$

We simplify now the rest term:

$$\frac{1}{2}\left(1-\frac{1}{\lambda}\right)^{-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\lambda\widetilde{\sigma}}\right)^2-$$

$$-\frac{1}{2}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2+\frac{1}{2\lambda}\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2$$

$$
\begin{aligned}
&= \frac{1}{2}\Bigg\{\left[\left(1-\frac{1}{\lambda}\right)^{-1}-1\right]\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2+ \\
&\quad+\left[\frac{1}{\lambda^2}\left(1-\frac{1}{\lambda}\right)^{-1}+\frac{1}{\lambda}\right]\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2- \\
&\quad-\frac{2}{\lambda}\left(1-\frac{1}{\lambda}\right)^{-1}\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\;\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\Bigg\} \\
&= \frac{1}{2}\left[\frac{1}{\lambda-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2+\frac{1}{\lambda-1}\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2-\right. \\
&\quad\left.-\frac{2}{\lambda-1}\;\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\;\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right] \\
&= \frac{1}{2(\lambda-1)}\left[\frac{\left(\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t\right)-\left(\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t\right)}{\widetilde{\sigma}}\right]^2 \\
&= \frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)}-\overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2.
\end{aligned}
$$

Putting all together we have

$$
\begin{aligned}
&\frac{\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}},\sqrt{\lambda}\right)(z)}= \\
&=\sqrt{\lambda}\exp\left[-\frac{1}{2}\left(z-\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2+\frac{1}{2\lambda}\left(z-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2\right] \\
&=\sqrt{\lambda}\exp\Bigg\{-\frac{1}{2}\left(1-\frac{1}{\lambda}\right)\left[z-\left(1-\frac{1}{\lambda}\right)^{-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\lambda\widetilde{\sigma}}\right)\right]^2+ \\
&\quad+\frac{1}{2}\left(1-\frac{1}{\lambda}\right)^{-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\lambda\widetilde{\sigma}}\right)^2- \\
&\quad-\frac{1}{2}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2+\frac{1}{2\lambda}\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}\right)^2\Bigg\}
\end{aligned}
$$

$$= \sqrt{\lambda} \exp\left[\frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)} - \overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2\right] \cdot$$

$$\cdot \exp\left\{-\frac{1}{2}\left(1-\frac{1}{\lambda}\right)\left[z-\left(1-\frac{1}{\lambda}\right)^{-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\lambda\widetilde{\sigma}}\right)\right]^2\right\}.$$

Now, the second assertion: the proof of the inequality follows directly from the first result. If $\lambda > 1$, then $\left(1 - \frac{1}{\lambda}\right) > 0$. Thus, the second exponential term in the first result has a negative exponent and it is at most 1. Moreover, the first derivative of

$$h(\lambda) = \sqrt{\lambda} \exp\left[\frac{e_j}{2(\lambda-1)}\right]$$

is

$$h'(\lambda) = \frac{1}{2}\left[\frac{1}{\sqrt{\lambda}} - \frac{\sqrt{\lambda}}{(\lambda-1)^2} e_j\right] \cdot \exp\left[\frac{e_j}{2(\lambda-1)}\right].$$

Then, $h'(\lambda) = 0$ if

$$\begin{aligned} 0 &= \left[\frac{1}{\sqrt{\lambda}} - \frac{\sqrt{\lambda}}{(\lambda-1)^2} e_j\right], \\ 0 &= (\lambda-1)^2 - \lambda e_j, \\ 0 &= \lambda^2 - (2+e_j)\lambda + 1. \end{aligned}$$

Therefore, the roots are

$$\begin{aligned} \lambda_{\pm} &= \frac{2+e_j \pm \sqrt{(2+e_j)^2 - 4}}{2} = \frac{2+e_j \pm \sqrt{e_j^2 + 4e_j}}{2} \\ &= \frac{4+2e_j \pm 2\sqrt{e_j}\sqrt{e_j+4}}{4} = \left(\frac{\sqrt{e_j} \pm \sqrt{e_j+4}}{2}\right)^2. \end{aligned}$$

We see from the sign of the first derivative that λ_- is a maximum and λ_+ is a minimum of the function $h(\lambda)$. □

Thus, we can construct the optimal distribution $(\tau_j(i))$ given the proposals $\rho(i, z)$.

Lemma 3.5 *Assume $\lambda > 1$ and independent of both z and the filter sample $(z^{(i)}_{(t-p):(t-1)})$.*
Then the optimal distribution $(\tau_j(i))$ is given by

$$\tau_j(i) = \frac{w_{i,j} \sum_{l=1}^{K} R_{l,i} M_{l,i}}{\sum_{i=1}^{N} w_{i,j} \sum_{l=1}^{K} R_{l,i} M_{l,i}}$$

with $w_{i,j}$ as in (3.9), $R_{k,i}$ and $M_{k,i}$ as in Lemma 3.3.

Proof: The proof is very similar to the corresponding one in the filtering case, see Lemma 2.4. It follows by applying Lemmas 2.1, 3.3 and 3.4. Note that the target density is given by $\widehat{p}\left(z|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$ in (3.10).
From Lemma 2.1, the optimal $\tau_j(i)$'s are given by

$$\tau_j(i) \ \propto M_i$$

with

$$M_i \geq \sup_z \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \ \ w_{i,j} \ \phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{\rho\left(i, z\right)}.$$

We find using the inequality (3.11) and Lemma 3.4 with $\lambda > 1$:

$$\begin{aligned}
&\frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \ \ w_{i,j} \ \phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{\rho\left(i, z\right)} \\
&= w_{i,j} \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \cdot \phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}, \sqrt{\lambda}\right)(z)}{\rho\left(i, z\right)} \ \frac{\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}, \sqrt{\lambda}\right)(z)} \\
&\leq w_{i,j} \ k_1 \ \left(\sum_{l=1}^{K} R_{l,i} M_{l,i}\right) \ \sqrt{\lambda} \ \exp\left[\frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)} - \overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2\right].
\end{aligned} \tag{3.12}$$

The last expression is independent of z, assuming that λ is independent of it. Thus, M_i can be set equal to this expression and the lemma follows since k_1, $\widetilde{\sigma}$ and $\mu_j^{(+)}$ are independent of the past sample

$(z^{(i)}_{(t-p):(t-1)})$ and the same is assumed for λ. □

The optimal distribution $(\tau_j(i))$ in Lemma 3.5 is similar to the one of the filtering case. In addition, there are the weights $w_{i,j}$ here. Unfortunately, these weights complicate the recursion since they also depend on the smoothing sample $(z^{(j)}_{(t+1):(t+p)})$. Thus, if we work directly with the distribution $(\tau_j(i))$ defined in the lemma, the $\tau_j(i)$'s have to be computed for each particle $z^{(j)}_{(t+1):(t+p)}$. In addition, their partial sums should also be evaluated because the sampling from $(\tau_j(i))$ is typically performed with the inversion method. All these computations slow down the algorithm.

We tried to eliminate the dependence on the smoothing sample as done in the construction of the proposals $\rho\,(i,z)$. I.e. we took weights

$$\overline{w}_{i,j} \propto \exp\left(-\frac{1}{2}\lambda_1 \left\|\overline{Rv^{(+)}_j} - Rv^{(-)}_i\right\|_2^2\right)$$

where $\overline{Rv^{(+)}_j}$ denotes the vector $Rv^{(+)}_j$ for a fixed particle $z^{(j)}_{(t+1):(t+p)}$. We had also to introduce λ_1 to normalise the acceptance probabilities. Then, we used these weights together with the previous proposals $\rho\,(i,z)$ to sample $z^{(j)}$. If the number of rejections exceeded a given bound, we stopped the sampling and we updated the distributions $\rho\,(i,z)$ and $(\tau(i))$. The new distributions were computed with the particle $z^{(j')}_{(t+1):(t+p)}$ for which the stop took place. This idea worked fine with a model where the state equation was given by an AR(1) or an AR(2) process. For AR processes with higher order, we got the well-known problem of the curse of dimensionality: the particles in a high dimensional space are sparse. Thus, the nearest neighbour $Rv^{(+)}_{j'}$ of $\overline{Rv^{(+)}_j}$ $(j' \neq j)$ was already too far and the weights $\overline{w}_{i,j}$ were a bad approximation of the optimal ones for j'. Consequently, we got a lot of rejections for each smoothing particle and many distribution setups were computed (for a moderate value of the maximal bound, say 500, a new distribution setup was computed for almost each smoothing particle $z^{(j)}_{(t+1):(t+p)}$). We also tried to separate the update of the distribution $(\tau(i))$ from the update of the proposals $\rho\,(i,z)$. The aim was to permit more updates of $(\tau(i))$ for the same proposals. But we got no significant improvements. In fact, if we compute new proposals $\rho\,(i,z)$, it is reasonable to find also the new distribution $(\tau(i))$ since the optimal $\tau(i)$

depends on the variables $R_{k,i}$ and $M_{k,i}$.
To overcome these difficulties, we have to think the sampling technique over. This gives the input to develop the second sampling method, see Section 3.4. For the first method, we are content with the procedure which uses the proposals $\rho\,(i,z)$ defined in Lemma 3.3 and it computes the distribution $(\tau_j(i))$ for each smoothing particle $z^{(j)}_{(t+1):(t+p)}$. In this way, we eliminate only the dependence on the smoothing sample in the construction of the proposals $\rho\,(i,z)$. We should expect that this algorithm is slow since it has a complexity of order TN^2.

The next step is to evaluate the acceptance probability of a proposed pair (i,z).

Lemma 3.6 *Assume $\lambda > 1$ and independent of both z and the filter sample $(z^{(i)}_{(t-p):(t-1)})$. Let the densities $\rho\,(i,z)$ and the distribution $(\tau_j(i))$ be defined as in Lemmas 3.3 and 3.5, respectively.*
Then the acceptance probability of the pair (i,z) generated from the distribution $\tau_j(i)\rho\,(i,z)$ is

$$\pi_j\,(i,z) = \exp\left\{-m\left[\log\left(1+\left(\frac{z}{c/\widetilde{\sigma}}\right)^2\right)+\alpha_{k^*}+\beta_{k^*}z+\gamma_{k^*}z^2\right]\right\}\cdot$$
$$\cdot\exp\left\{-\frac{1}{2}\Big(1-\frac{1}{\lambda}\Big)\left[z-\Big(1-\frac{1}{\lambda}\Big)^{-1}\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}-\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\lambda\widetilde{\sigma}}\right)\right]^2\right\}$$

where B_{k^*} is the subset containing z.

Remark 3.3 *The resulting acceptance probability has a very intuitive form as in the filtering case, see Lemma 2.5. It reflects the two approximations used to implement this method: the approximation of the logarithm of the Pearson type VII density (up to some terms) by a majorant and the approximation due to the substitution of $\mu_j^{(+)}$ by $\overline{\mu_j^{(+)}}$. Therefore, it is again important that the majorant approximates well the logarithm of the Pearson type VII density (up to some terms) to have a high acceptance probability.*

Proof: The proof is similar to the corresponding one in the filtering case, see Lemma 2.5. We take the formula for the acceptance probability in the rejection method with the auxiliary index and we evaluate it using the proposal densities $\rho\,(i,z)$ and the distribution $(\tau_j(i))$ (see their definitions in Lemmas 3.3 and 3.5). In addition, the result of Lemma 3.4 is used. Note that the target density is given by $\widehat{p}\left(z|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$ in (3.10).

The acceptance probability is given equivalently to (2.16) by

$$\pi_j\left(i,z\right) = \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \;\; w_{i,j} \; \phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\tau_j(i) \;\rho\left(i,z\right) \;\; M}$$

with

$$M \geq \sup_{i,z} \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \;\; w_{i,j} \; \phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\tau_j(i) \;\rho\left(i,z\right)}.$$

First, we calculate the supremum term and we define M. This can be achieved using again the inequality (3.11), the definition of the distribution $(\tau_j(i))$ in Lemma 3.5 and Lemma 3.4 with $\lambda > 1$.

$$\frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \;\; w_{i,j} \; \phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\tau_j(i) \;\rho\left(i,z\right)}$$

$$= \frac{w_{i,j}}{\tau_j(i)} \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \cdot \phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}},\sqrt{\lambda}\right)(z)}{\rho\left(i,z\right)} \; \frac{\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}},\sqrt{\lambda}\right)(z)}$$

$$\leq \frac{w_{i,j} \sum_{i=1}^N w_{i,j} \sum_{l=1}^K R_{l,i} M_{l,i}}{w_{i,j} \sum_{l=1}^K R_{l,i} M_{l,i}} \cdot k_1 \cdot \left(\sum_{l=1}^K R_{l,i} M_{l,i}\right) \cdot \frac{\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}},\sqrt{\lambda}\right)(z)}$$

$$\leq k_1 \left(\sum_{i=1}^N w_{i,j} \sum_{l=1}^K R_{l,i} M_{l,i}\right) \sqrt{\lambda} \; \exp\left[\frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)} - \overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2\right]. \tag{3.13}$$

The last expression does not depend neither on z nor on the filter sample $(z^{(i)}_{(t-p):(t-1)})$ for the assumed λ. Therefore, M can be set equal to it.

Using the definitions of M and $(\tau_j(i))$, the acceptance probability is given by

$$\pi_j(i,z) = \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z)\ \phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{k_1\left(\sum_{l=1}^{K} R_{l,i}M_{l,i}\right)\sqrt{\lambda}\ \exp\left[\frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)}-\overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2\right]\ \rho(i,z)}$$

$$= \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z)\cdot\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}, \sqrt{\lambda}\right)(z)}{k_1\cdot\left(\sum_{l=1}^{K} R_{l,i}M_{l,i}\right)\cdot\rho(i,z)} \frac{\frac{\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}, \sqrt{\lambda}\right)(z)}}{\sqrt{\lambda}\exp\left[\frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)}-\overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2\right]}. \tag{3.14}$$

Now, the first fraction can be simplified as in the filtering case, see Lemma 2.5, equation (2.24). Substitute again σ_V by $\widetilde{\sigma}$, $\sum_{l=1}^{p}\varphi_l z_{t-l}^{(i)}$ by $\mu_i^{(-)}+\overline{\mu_j^{(+)}}$ and recall that z is in B_{k^*}. The second fraction can be evaluated using Lemma 3.4. Thus, the lemma follows. □

There are still two topics to discuss: the choice of λ and the construction of the majorant.

What is a suitable value for λ? λ should be computable at the beginning of each smoothing step since it is needed to construct the proposals $\rho(i,z)$ and the distribution $(\tau_j(i))$. Up to now, we have required that λ is greater than one and that it is independent of z, the filter sample and the smoothing sample. In addition, we look for a simple procedure to compute λ. Thus, an intuitive strategy is to minimize the right side of the inequality

$$\frac{\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{\phi\left(\frac{\mu_i^{(-)}+\overline{\mu_j^{(+)}}-\widetilde{y}_t}{\widetilde{\sigma}}, \sqrt{\lambda}\right)(z)} \leq \sqrt{\lambda}\ \exp\left[\frac{1}{2(\lambda-1)}\left(\frac{\mu_j^{(+)}-\overline{\mu_j^{(+)}}}{\widetilde{\sigma}}\right)^2\right]$$

with respect to λ, see Lemma 3.4. In this way, we will improve the approximations in the construction of the distribution $(\tau_j(i))$ and in the evaluation of the acceptance probability and also increase the acceptance probability, see (3.12), (3.13) and (3.14). Contemporaneously, we

should consider the mentioned restrictions.
Explicitly, Lemma 3.4 gives the value of λ at which the minimum of the right side is reached. We propose the following procedure to find a suitable value of λ given all $\mu_j^{(+)}$:

- Compute $\widetilde{e}_j := \left(\mu_j^{(+)} - med_j\left(\mu_j^{(+)}\right)\right)^2$, for $j = 1, \ldots, N$.
- Compute $\widetilde{e} := \frac{1}{\widetilde{\sigma}^2}\ med_j\left(\widetilde{e}_j\right)$.
- Set $\lambda_{tmp} := \left(\frac{\sqrt{\widetilde{e}}+\sqrt{\widetilde{e}+4}}{2}\right)^2$. Note that λ_{tmp} is at least one.
- The experience shows that more than one distribution setup is computed to generate the sample $(z^{(j)})$. Thus, it is reasonable to reduce the previous defined value of λ. We take

  $$\lambda := 1 + \frac{\lambda_{tmp} - 1}{2} = \frac{1}{2} + \frac{1}{2}\left(\frac{\sqrt{\widetilde{e}} + \sqrt{\widetilde{e}+4}}{2}\right)^2 .$$

  In the (unrealistic) case where the resulting λ is one, we set $\lambda = 1.01$.

This choice of λ fulfils the restrictions discussed before.

In Subsection 2.3.4, we discussed the construction of the majorant in the filtering recursion. We found out that it was reasonable to distinguish three cases with corresponding majorants. The same arguments are also valid here. Then, we state equivalently to the filtering case:

Definition 3.2 *Let the $\mu_i^{(-)}$'s, $\overline{\mu_j^{(+)}}$, $\widetilde{\sigma}$ and λ be as in Lemma 3.3 and let δ be a strictly positive real number. Define*

$$\mu := med_i\left(\frac{\mu_i^{(-)} + \overline{\mu_j^{(+)}} - \widetilde{y}_t}{\widetilde{\sigma}}\right),$$

$$l_{(-)} := g(\mu - \sqrt{\lambda}\delta) = -\log\left[1 + \left(\frac{\mu - \sqrt{\lambda}\delta}{c/\widetilde{\sigma}}\right)^2\right],$$

$$l_{(+)} := g(\mu + \sqrt{\lambda}\delta) = -\log\left[1 + \left(\frac{\mu + \sqrt{\lambda}\delta}{c/\widetilde{\sigma}}\right)^2\right].$$

Then, the parameters to define the default majorant are given in Table 3.1, the parameters for the lower majorant in Table 3.2 and the *parameters for the upper majorant in Table 3.3. Note that lower and upper majorants can be defined only if* $\mu < -\sqrt{3}\ c/\widetilde{\sigma} - \sqrt{\lambda}\delta$ *and* $\mu > \sqrt{3}\ c/\widetilde{\sigma} + \sqrt{\lambda}\delta$*, respectively.*

k	d_{k-1}	d_k	α_k	β_k	γ_k
1	$-\infty$	$-5\ c/\widetilde{\sigma}$	$-\log(26)$	0	0
2	$-5\ c/\widetilde{\sigma}$	$-\sqrt{3}\ c/\widetilde{\sigma}$	$\frac{\sqrt{3}\log(6.5)}{5-\sqrt{3}} - \log(4)$	$\frac{\log(6.5)}{\left(5-\sqrt{3}\right)\ c/\widetilde{\sigma}}$	0
3	$-\sqrt{3}\ c/\widetilde{\sigma}$	$-c/\widetilde{\sigma}$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\widetilde{\sigma}}$	0
4	$-c/\widetilde{\sigma}$	$c/\widetilde{\sigma}$	0	0	$-\frac{\log(2)}{(c/\widetilde{\sigma})^2}$
5	$c/\widetilde{\sigma}$	$\sqrt{3}\ c/\widetilde{\sigma}$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$-\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\widetilde{\sigma}}$	0
6	$\sqrt{3}\ c/\widetilde{\sigma}$	$5\ c/\widetilde{\sigma}$	$\frac{\sqrt{3}\log(6.5)}{5-\sqrt{3}} - \log(4)$	$-\frac{\log(6.5)}{\left(5-\sqrt{3}\right)\ c/\widetilde{\sigma}}$	0
7	$5\ c/\widetilde{\sigma}$	∞	$-\log(26)$	0	0

Table 3.1: *Parameters to define the default majorant in the smoothing recursion (first method).*

k	d_{k-1}	d_k	α_k	β_k	γ_k
1	$-\infty$	$\mu - \sqrt{\lambda}\delta$	$l_{(-)}$	0	0
2	$\mu - \sqrt{\lambda}\delta$	$\mu + \sqrt{\lambda}\delta$	$l_{(-)} - \beta_2\ (\mu - \sqrt{\lambda}\delta)$	$\frac{l_{(+)}-l_{(-)}}{2\sqrt{\lambda}\delta}$	0
3	$\mu + \sqrt{\lambda}\delta$	$-\sqrt{3}\ c/\widetilde{\sigma}$	$l_{(+)} - \beta_3\ (\mu + \sqrt{\lambda}\delta)$	$\frac{\log(4)+l_{(+)}}{\mu+\sqrt{\lambda}\delta+\sqrt{3}\ c/\widetilde{\sigma}}$	0
4	$-\sqrt{3}\ c/\widetilde{\sigma}$	$-c/\widetilde{\sigma}$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\widetilde{\sigma}}$	0
5	$-c/\widetilde{\sigma}$	$c/\widetilde{\sigma}$	0	0	$-\frac{\log(2)}{(c/\widetilde{\sigma})^2}$
6	$c/\widetilde{\sigma}$	$\sqrt{3}\ c/\widetilde{\sigma}$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$-\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\widetilde{\sigma}}$	0
7	$\sqrt{3}\ c/\widetilde{\sigma}$	∞	$-\log(4)$	0	0

Table 3.2: *Parameters to define the lower majorant in the smoothing recursion (first method). We should have* $\mu < -\sqrt{3}\ c/\widetilde{\sigma} - \sqrt{\lambda}\delta$.

We select:

- *the default majorant if* $-\sqrt{3}\ c/\widetilde{\sigma} - \sqrt{\lambda}\delta \leq \mu \leq \sqrt{3}\ c/\widetilde{\sigma} + \sqrt{\lambda}\delta$,
- *the lower majorant if* $-\sqrt{3}\ c/\widetilde{\sigma} - \sqrt{\lambda}\delta > \mu$,
- *the upper majorant if* $\mu > \sqrt{3}\ c/\widetilde{\sigma} + \sqrt{\lambda}\delta$.

k	d_{k-1}	d_k	α_k	β_k	γ_k
1	$-\infty$	$-\sqrt{3}\ c/\widetilde{\sigma}$	$-\log(4)$	0	0
2	$-\sqrt{3}\ c/\widetilde{\sigma}$	$-c/\widetilde{\sigma}$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\widetilde{\sigma}}$	0
3	$-c/\widetilde{\sigma}$	$c/\widetilde{\sigma}$	0	0	$-\frac{\log(2)}{(c/\widetilde{\sigma})^2}$
4	$c/\widetilde{\sigma}$	$\sqrt{3}\ c/\widetilde{\sigma}$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$-\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\widetilde{\sigma}}$	0
5	$\sqrt{3}\ c/\widetilde{\sigma}$	$\mu-\sqrt{\lambda\delta}$	$-\log(4)-\beta_5\ \sqrt{3}\ c/\widetilde{\sigma}$	$\frac{l_{(-)}+\log(4)}{\mu-\sqrt{\lambda\delta}-\sqrt{3}\ c/\widetilde{\sigma}}$	0
6	$\mu-\sqrt{\lambda\delta}$	$\mu+\sqrt{\lambda\delta}$	$l_{(-)}-\beta_6\ (\mu-\sqrt{\lambda\delta})$	$\frac{l_{(+)}-l_{(-)}}{2\sqrt{\lambda\delta}}$	0
7	$\mu+\sqrt{\lambda\delta}$	∞	$l_{(+)}$	0	0

Table 3.3: *Parameters to define the upper majorant in the smoothing recursion (first method). We should have* $\mu > \sqrt{3}\ c/\widetilde{\sigma} + \sqrt{\lambda\delta}$.

The remarks done in the filtering case are also valid here. Moreover:

Lemma 3.7 *The three majorants in Definition 3.2 satisfy Definition 3.1. I.e., they fulfil the inequality*

$$g(z) = -\log\left[1+\left(\frac{z}{c/\widetilde{\sigma}}\right)^2\right] \le \sum_{k=1}^{K}\left(\alpha_k+\beta_k z+\gamma_k z^2\right)\mathbb{I}_{\{z\in B_k\}}, \ \forall z\in\mathbb{R}.$$

In addition, these majorants are continuous functions in z.

Proof: The proof is the same as in the filtering case, see Lemma 2.6. Substitute σ_V by $\widetilde{\sigma}$. □

Finally, we put together all previous results and we summarize the particle smoothing recursion at time t with $\widetilde{y}_t$ known.

Algorithm 3.1 *Particle smoothing recursion at time* t *with* $\widetilde{y}_t$ *available.*

Assumptions:

- *The fully defined default majorant and the partially defined lower and upper majorants have already been computed and* δ *is known, see Definition 3.2.*

- *The upper triangular matrix R has already been found as explained in Remark 3.1.*
- *The maximal allowed number of rejections is known.*
- *The filter sample $(z^{(i)}_{(t-p):(t-1)})$ and the smoothing particles $(z^{(j)}_{(t+1):(t+p)})$, $i, j = 1, \dots, N$, are known from the filtering recursion and the previous smoothing step at time $t+1$, respectively.*

Preliminaries:

1. *Compute the $\mu_i^{(-)}$'s and the $\mu_j^{(+)}$'s for $i, j = 1, \dots, N$ as described in (3.5), (3.6) and Lemma 3.1.*
   Compute the $Rv_i^{(-)}$'s and the $Rv_j^{(+)}$'s for $i, j = 1, \dots, N$ with R as above and $v_i^{(-)}$'s, $v_j^{(+)}$'s as in (3.7), (3.8) and Lemma 3.1.
   Sort the $\mu_j^{(+)}$'s according to size and apply the same permutation to the $Rv_j^{(+)}$'s.
2. *Compute λ.*

   To this end:

   (a) *Compute $\widetilde{e}_j := \left(\mu_j^{(+)} - med_j\left(\mu_j^{(+)}\right)\right)^2$, for $j = 1, \dots, N$.*
   (b) *Compute $\widetilde{e} := \frac{1}{\widetilde{\sigma}^2}\, med_j\,(\widetilde{e}_j)$.*
   (c) *Set $\lambda_{tmp} := \left(\frac{\sqrt{\widetilde{e}}+\sqrt{\widetilde{e}+4}}{2}\right)^2$.*
   (d) *Set*

   $$\lambda := 1 + \frac{\lambda_{tmp} - 1}{2} = \frac{1}{2} + \frac{1}{2}\left(\frac{\sqrt{\widetilde{e}} + \sqrt{\widetilde{e}+4}}{2}\right)^2 .$$

   In the (unrealistic) case where the resulting λ is 1, set $\lambda = 1.01$.

Begin the construction of the particles $(z_t^{(j)})$, $j = 1, \dots, N$. Set $j = 1$.

3. *Set $\overline{\mu_j^{(+)}} = \mu_j^{(+)}$ and choose the majorant.*

   For this purpose, define

   $$\mu = med_i \left(\frac{\mu_i^{(-)} + \overline{\mu_j^{(+)}} - \widetilde{y}_t}{\widetilde{\sigma}} \right)$$

and select

- *the default majorant if* $-\sqrt{3}\,\frac{c}{\tilde{\sigma}} - \sqrt{\lambda}\delta \leq \mu \leq \sqrt{3}\,\frac{c}{\tilde{\sigma}} + \sqrt{\lambda}\delta$,
- *the lower majorant if* $-\sqrt{3}\,\frac{c}{\tilde{\sigma}} - \sqrt{\lambda}\delta > \mu$,
- *the upper majorant if* $\mu > \sqrt{3}\,\frac{c}{\tilde{\sigma}} + \sqrt{\lambda}\delta$.

If the lower or the upper majorant is chosen, compute the components which depend on μ, λ *and* δ, *see Definition 3.2.*

4. *Compute the distribution setup for the rejection sampling method using* $\overline{\mu_j^{(+)}}$. *I.e. find the variables to get the efficient proposals* $\rho\,(i,z)$ *and the optimal distribution* $(\tau_j(i))$.

   To this end, compute the variables $\overline{\sigma}_k$, $\overline{\mu}_{k,i}$, $\Phi\left(\frac{d_k-\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)$, $\Phi\left(\frac{d_{k-1}-\overline{\mu}_{k,i}}{\overline{\sigma}_k}\right)$ *and* $RM_{k,i} := R_{k,i}\cdot M_{k,i}$ *for all* $k=1,\ldots,K$ *and all* $i=1,\ldots,N$ *as described in Lemma 3.3. In addition, compute the partial sums of the* $RM_{k,i}$*'s over* k *for all* i.

5. *Compute the optimal distribution* $(\tau_j(i))$ *for the particle* $z^{(j)}_{(t+1):(t+p)}$ *according to Lemma 3.5. Find also its partial sums. In addition, set the rejection counter* rej-counter *to* 0.

6. *Sample a pair* $\left(i^{(j)},z^{(j)}\right)$ *according to the distribution* $\tau_j(i)\rho\,(i,z)$.

   First, generate the auxiliary index $I^{(j)}$ *according to the distribution* $(\tau_j(i))$ *and then the variable* $Z^{(j)}$ *according to the density* $\rho\left(i^{(j)},z\right)$ *with* $I^{(j)}=i^{(j)}$. *The two samplings are carried out by the inversion method, see also Remark 2.6.*

7. *Check the acceptance of the proposed pair* $\left(i^{(j)},z^{(j)}\right)$.

   To this end, generate U *uniform on* $[0,1]$ *and compute the acceptance probability* $\pi_j\left(i^{(j)},z^{(j)}\right)$ *according to Lemma 3.6.*

   - *If* $U\leq\pi_j\left(i^{(j)},z^{(j)}\right)$, *then accept the pair* $\left(i^{(j)},z^{(j)}\right)$. *Return the particle* $z_t^{(j)}$ *defined by*

     $$z_t^{(j)} = \widetilde{y}_t + \widetilde{\sigma}\; z^{(j)}.$$

     Set $j=j+1$.
     If $j\leq N$, *return to step 5. Otherwise stop: all particles* $(z_t^{(j)})$ *have been computed.*

   - *Else, the pair* $\left(i^{(j)},z^{(j)}\right)$ *is not accepted. Increment* rej-counter *by* 1.

> *If* rej-counter $\leq$ *maximal allowed number of rejections, return to step 6. Else, return to step 3 (improve the majorant and then the setup).*

3.3.2 Recursion with missing $\widetilde{y}_t$

We discuss now the smoothing recursion when $\widetilde{y}_t$ is missing. The recursion simplifies considerably as in the filtering case, see Section 2.4.

If $\widetilde{y}_t$ is not available, we do not have the Bayes' step (3.1) in the exact computation of the density $p\left(z_t|z_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$. Thus, for each $z^{(j)}_{(t+1):(t+p)}$ we should generate a particle $z^{(j)}_t$ from the density $\widehat{p}\left(z_t|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$ defined by

$$\widehat{p}\left(z_t|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right) = \sum_{i=1}^{N} \frac{w_{i,j}}{\sum_{i=1}^{N} w_{i,j}} \phi\left(\mu^{(-)}_i + \mu^{(+)}_j, \widetilde{\sigma}\right)(z_t), \quad (3.15)$$

see (3.3) and (3.4). Note that we have not dropped $\widetilde{y}_t$ from the previous smoothing density notation, although it is missing. We have set the "value" of $\widetilde{y}_t$ to *NA* (*not available*). In this way, the formula has a better readability.

The density (3.15) is a mixture of normal densities with weights given by $w_{i,j}/\sum_{i=1}^{N} w_{i,j}$. Both the normal densities and the weights depend on the filter and smoothing samples. The latter dependence causes the same difficulties met in the recursion with $\widetilde{y}_t$ available. Then, we use a two-step procedure to sample from this mixture: we generate the index I according to the weight distribution and then the variable Z_t is sampled from the normal density $\phi\left(\mu^{(-)}_i + \mu^{(+)}_j, \widetilde{\sigma}\right)(z_t)$ with $I = i$. The sampling from the normal density is straightforward if we take care to compute all $\mu^{(-)}_i$'s and all $\mu^{(+)}_j$'s at the beginning of the smoothing step. The difficulty is given by the sampling from the weight distribution. This sampling is achieved typically with the inversion method and thus the evaluation of the partial sums is also required. But, since the weights depend on the smoothing sample, the weight distribution and its partial sums have to be computed for each smoothing particle. Thus, the resulting algorithm is not particularly fast. In Subsection 3.3.1, we discussed some ideas to avoid that the weights depended on the smoothing particles. But unfortunately, we did not come off. We have to think the sampling procedure over to have an effective improvement, see Section 3.4. Again, for method 1, we are satisfied with the described two-step sampling method.

Then, the smoothing recursion can be summarized as follows.

Algorithm 3.2 *Particle smoothing recursion at time t with missing $\widetilde{y}_t$.*
Assumptions:

- *The upper triangular matrix R has already been found as explained in Remark 3.1.*
- *The filter sample $(z^{(i)}_{(t-p):(t-1)})$ and the smoothing particles $(z^{(j)}_{(t+1):(t+p)})$, $i, j = 1, \ldots, N$, are known from the filtering recursion and the previous smoothing step at time $t+1$, respectively.*

Preliminaries:

1. *Compute the $\mu^{(-)}_i$'s and the $\mu^{(+)}_j$'s for $i, j = 1, \ldots, N$ as described in (3.5), (3.6) and Lemma 3.1.*
   Compute the $Rv^{(-)}_i$'s and the $Rv^{(+)}_j$'s for $i, j = 1, \ldots, N$ with R as above and $v^{(-)}_i$'s, $v^{(+)}_j$'s as in (3.7), (3.8) and Lemma 3.1.

Begin the construction of the particles $(z^{(j)}_t)$, $j = 1, \ldots, N$.

2. *For j from 1 to N do:*
   - *(a) Compute the weights*
   $$\frac{w_{i,j}}{\sum_{i=1}^{N} w_{i,j}}, \quad i = 1, \ldots, N$$
   with $w_{i,j}$ as in (3.9). In addition, compute the partial sums of the weights.
   - *(b) Sample $I^{(j)}$ according to the weight distribution using the inversion method.*
   - *(c) Sample $Z^{(j)}_t$ from $\phi\left(\mu^{(-)}_{i^{(j)}} + \mu^{(+)}_j, \widetilde{\sigma}\right)(z_t)$ with $I^{(j)} = i^{(j)}$. Return $z^{(j)}_t$.*

3.4 Particle smoothing recursion: method 2

In Section 3.3, we discussed the first method to implement the smoothing recursion at time t. It was derived from the filtering one but the resulting algorithm was not particularly fast. It is useful to sketch again the leading ideas to have a starting point for the development of the second method.

The task is to generate a particle $z_t^{(j)}$ from the density $\widehat{p}\left(z_t|z_{(t+1):(t+p)}^{(j)}, \widetilde{y}_{1:t}\right)$ for each $z_{(t+1):(t+p)}^{(j)}$. Thanks to Lemma 3.2, we can generate a particle $z^{(j)}$ from each density $\widehat{p}\left(z|z_{(t+1):(t+p)}^{(j)}, \widetilde{y}_{1:t}\right)$ defined as in (3.10). In the first method, this sampling was attained with the rejection method with the auxiliary index. Unfortunately, we saw that the efficient proposal densities $\rho_j(i,z)$ depended on both the filter sample $(z_{(t-p):(t-1)}^{(i)})$ and the smoothing particles $(z_{(t+1):(t+p)}^{(j)})$. The latter dependence is harmful and it was eliminated by constructing the proposal densities $\rho(i,z)$ using a fixed smoothing particle $z_{(t+1):(t+p)}^{(j)}$ and, consequently, by enlarging the variance of the normal densities. In addition, an update of these proposal densities was done once they became bad proposals. Although the proposals $\rho(i,z)$ were used to sample several $z^{(j)}$'s, this approach was not very efficient because the proposals were computed for all filter particles $(z_{(t-p):(t-1)}^{(i)})$. This required a huge amount of computations. The computation of the auxiliary distribution $\tau_j(i)$ was even worse. In fact the optimal choice depended again on both filter and smoothing particles and we did not succeed in eliminating the second dependence in a clever way. At that time, we proposed to compute the auxiliary distribution $\tau_j(i)$ and its partial sums for each smoothing particle $z_{(t+1):(t+p)}^{(j)}$.

How can we speed up the algorithm? The key concept is to compute a distribution setup which is simultaneously reliable and contains few distributions since only one particle $z^{(j)}$ is sampled from the target density $\widehat{p}\left(z|z_{(t+1):(t+p)}^{(j)}, \widetilde{y}_{1:t}\right)$. Thus, we compute again the auxiliary distribution $(\tau_j(i))$ for each j but we take care to group similar filter particles and we give the same value to the corresponding $\tau_j(i)$'s. In addition, we can introduce a pretesting on the sampled index I and we carry out the construction of the proposal density $\rho_j(i,z)$ for this index I only if it has passed the pretesting. In this way, an unfavourable index I is discarded immediately and it does not cause additional computations. Thus, we can allow that the auxiliary distribution and the proposal density depend on the smoothing particle $z_{(t+1):(t+p)}^{(j)}$ since we reduce the amount of required computations to find the auxiliary distribution and we compute the proposal density only when it is needed.

In few words, these are the leading ideas to develop the second method. We discuss them in detail in the following subsections. We distinguish again between the case with $\widetilde{y}_t$ available and the one with missing $\widetilde{y}_t$. Since the latter is a special case of the former, we discuss first the recursion with $\widetilde{y}_t$ available.

3.4.1 Recursion with $\widetilde{y}_t$ available

The aim is to improve the sampling of $z^{(j)}$ from the target density $\widehat{p}\left(z|z^{(j)}_{(t+1):(t+p)}, \widetilde{y}_{1:t}\right)$ with known $\widetilde{y}_t$. The sampling is performed again using the rejection method with the auxiliary index. Thus, there are several points to treat. We should find efficient distributions $\rho_j\left(i,z\right)$ and $\left(\tau_j(i)\right)$ and evaluate the acceptance probabilities of the proposed pairs (i,z). But we wish to apply the rejection method using a distribution setup that is reliable and contains only few distributions.

Let us begin with the construction of the proposal densities $\rho_j\left(i,z\right)$. They should be good proposal distributions for the densities proportional to

$$p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z) \cdot \phi\left(\frac{\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)$$

as seen in Subsection 2.2.1 and using that the target density is given by (3.10). Unlike the first method, we work now directly with the previous densities. I.e., we do not substitute $\mu_j^{(+)}$ by $\overline{\mu_j^{(+)}}$ and we do not enlarge the variance of the normal densities, since the construction of the densities $\rho_j\left(i,z\right)$ will be carried out only after the pretesting on the indices I, i.e. only for potentially good indices I. This is a first step towards having few distributions in the setup since $\rho_j\left(i,z\right)$ will be found for one index I at a time and not for all filter particles. The dependence on the smoothing particle $z^{(j)}_{(t+1):(t+p)}$ is allowed since it does not cause additional difficulties.

The majorant is defined in the same way as in the first method, see Definition 3.1. Again, we delay its construction and we concentrate on finding the proposals $\rho_j\left(i,z\right)$.

Lemma 3.8 *Let the proposal densities* $\rho_j\,(i,z)$ *be chosen as*

$$\rho_j\,(i,z) = \sum_{k=1}^{K} \frac{R_{k,ij} M_{k,ij}}{\sum_{l=1}^{K} R_{l,ij} M_{l,ij}} \; \frac{\exp\left[-\frac{1}{2}\left(\frac{z-\overline{\mu}_{k,ij}}{\overline{\sigma}_k}\right)^2\right]}{M_{k,ij}} \; \mathbb{I}_{\{z \in B_k\}}$$

with

$$\begin{aligned}
\overline{\sigma}_k &:= \sqrt{\frac{1}{1-2m\gamma_k}}, \\
\overline{\mu}_{k,ij} &:= \overline{\sigma}_k^2 \left(\frac{\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t}{\widetilde{\sigma}} + m\beta_k\right), \\
R_{k,ij} &:= \exp\left[\frac{1}{2}\left(\frac{\overline{\mu}_{k,ij}}{\overline{\sigma}_k}\right)^2 - \frac{1}{2}\left(\frac{\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t}{\widetilde{\sigma}}\right)^2 + m\alpha_k\right], \\
M_{k,ij} &:= \sqrt{2\pi}\;\overline{\sigma}_k \left[\Phi\left(\frac{d_k - \overline{\mu}_{k,ij}}{\overline{\sigma}_k}\right) - \Phi\left(\frac{d_{k-1} - \overline{\mu}_{k,ij}}{\overline{\sigma}_k}\right)\right].
\end{aligned}$$

The parameters α_k*,* β_k*,* γ_k *and* d_k *are as in Definition 3.1;* $\widetilde{\sigma}$ *is as in Lemma 3.1;* $\mu_i^{(-)}$ *and* $\mu_j^{(+)}$ *are as in (3.5) and (3.6);* m *and* c *are the parameters of the observation error distribution in the considered model (2.6) and (2.7). Finally,* $\Phi\,(x)$ *is the cumulative* $\mathcal{N}(0,1)$ *distribution function.*
Then

$$p_{VII}\,(m, c/\widetilde{\sigma}, 0)\,(z)\cdot\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z) \le k_1\cdot\left(\sum_{l=1}^{K} R_{l,ij} M_{l,ij}\right)\cdot\rho_j\,(i,z) \tag{3.16}$$

with

$$k_1 := \frac{\Gamma(m)\;\widetilde{\sigma}}{\sqrt{2}\;\pi\;c\;\Gamma(m-0.5)}.$$

Remark 3.4 *Most remarks of the corresponding filtering Lemma 2.3 are also valid here.*

Proof: The proof is again the same as in the filtering case, see Lemma 2.3. Substitute σ_V by $\widetilde{\sigma}$, $\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)}$ by $\mu_i^{(-)} + \mu_j^{(+)}$ and μ_i by $\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}$. In addition, set $\lambda = 1$ and use the new definitions of $\overline{\sigma}_k$,

$\overline{\mu}_{k,ij}$, $R_{k,ij}$ and k_1 given here. □

In Remark 2.6, we also gave a second interpretation of the inequality (3.16) and we found a useful feature of it. We generalize this characteristic as follows.

Definition 3.3 *Let $q(z)$ be a majorant of the density $p_{VII}(m, c/\widetilde{\sigma}, 0)(z)$, see also Remark 2.6. Then, we define*

$$M(a) := \widetilde{M}\left(\frac{a}{\widetilde{\sigma}}\right) := \frac{1}{k_1}\int q(z)\ \phi_1\left(z - \frac{a}{\widetilde{\sigma}}\right) dz.$$

Remark 3.5 *Some comments about the previous definition.*

- *In words, $\widetilde{M}\left(\frac{a}{\widetilde{\sigma}}\right)$ is proportional to the convolution of the two functions $q(z)$ and $\phi_1(z)$. Since $\widetilde{\sigma}$ depends only on the hyperparameters, we drop it from the notation and we denote $\widetilde{M}\left(\frac{a}{\widetilde{\sigma}}\right)$ simply by $M(a)$.*
- *If $\log(q(z))$ is defined using the Definition 3.1, then we have*

$$M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) = \sum_{l=1}^{K} R_{l,ij} M_{l,ij},$$

  see the inequality (3.16).
- *If $q(z)$ is symmetric about zero, the same is also true for $M(a)$. In fact, we get with the substitution $z' = -z$ and the symmetry of $\phi_1(z)$ about zero that:*

$$\begin{aligned} M(-a) &= \frac{1}{k_1}\int_{-\infty}^{\infty} q(z)\ \phi_1\left(z + \frac{a}{\widetilde{\sigma}}\right) dz \\ &= \frac{1}{k_1}\int_{\infty}^{-\infty} q(-z')\ \phi_1\left(-z' + \frac{a}{\widetilde{\sigma}}\right)(-dz') \\ &= \frac{1}{k_1}\int_{-\infty}^{\infty} q(z')\ \phi_1\left(z' - \frac{a}{\widetilde{\sigma}}\right) dz' = M(a). \end{aligned}$$

Since we have chosen the proposal densities $\rho_j(i, z)$, we can compute the optimal distribution $(\tau_j(i))$. We find

Lemma 3.9 *The optimal distribution $(\tau_j(i))$ is given by*

$$\tau_j(i) \;\propto\; w_{i,j}\; M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right)$$

with $\mu_i^{(-)}$, $\mu_j^{(+)}$ and $w_{i,j}$ as in (3.5), (3.6) and (3.9), respectively; $M\left(a\right)$ as in Definition 3.3.

Proof: The proof is very similar to the ones in the filtering case and in the smoothing recursion implemented with the first method, see Lemmas 2.4 and 3.5. I.e. it follows using Lemma 2.1, the inequality (3.16) and Remark 3.5. Note that the target density is given by $\widehat{p}\left(z|z_{(t+1):(t+p)}^{(j)}, \widetilde{y}_{1:t}\right)$, see (3.10).
From Lemma 2.1, the optimal $\tau_j(i)$'s are given by

$$\tau_j(i) \propto M_i$$

with

$$M_i \geq \sup_z \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z)\;\; w_{i,j}\;\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{\rho_j\left(i, z\right)}.$$

Using the inequality (3.16) and Remark 3.5, we find

$$\frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z)\;\; w_{i,j}\;\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{\rho_j\left(i, z\right)}$$

$$\leq k_1\; w_{i,j}\;\sum_{l=1}^{K} R_{l,ij} M_{l,ij} = k_1\; w_{i,j}\; M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) =: M_i$$

since the right hand side of the inequality is independent of z. The lemma follows because k_1 does not depend on the filter sample $(z_{(t-p):(t-1)}^{(i)})$. □

The optimal distribution $(\tau_j(i))$ depends on both the filter sample $(z_{(t-p):(t-1)}^{(i)})$ and the smoothing particles $(z_{(t+1):(t+p)}^{(j)})$ as we have already noted in the first smoothing method. At that time, we tried to eliminate the latter dependence, but without success. Now, the first key idea is to reduce the computation complexity with respect to the filter sample. This can be achieved with the following definition.

Definition 3.4 *The range*

$$\left[\min\left(\mu_i^{(-)}\right), \max\left(\mu_i^{(-)}\right)\right]$$

is divided in equally wide subsets $S_1, \ldots, S_{N_{t-1}}$. The index subsets I_r are defined by

$$I_r = \left\{ i \middle| \mu_i^{(-)} \in S_r \right\}, \qquad r = 1, \ldots, N_{t-1}.$$

Remark 3.6 *Since the $\mu_i^{(-)}$'s are constructed from the filter sample $(z_{(t-p):(t-1)}^{(i)})$ and the subset width is fixed, the number of subsets depends on this filter sample. We denote the dependence by N_{t-1}. In addition, note that some subsets I_r may be empty.*

We can combine the result of Lemma 3.9 with Definition 3.4.

Definition 3.5 *The used distribution $(\tau_j(i))$ is defined by*

$$\tau_j(i) = \tau_j(i(r)) = \frac{c_j(r)}{\sum_{r=1}^{N_{t-1}} |I_r| \, c_j(r)}$$

with $i \in I_r$, $r \in \{1, \ldots, N_{t-1}\}$. $c_j(r)$ is defined by

$$c_j(r) \geq \sup_{i \in I_r} \left(w_{i,j} \; M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) \right).$$

In words, the idea of the previous definition is to get the same value to all $\tau_j(i)$ with i in I_r. This reduces the computation complexity since we have to compute only N_{t-1} values $c_j(r)$ to define the distribution $(\tau_j(i))$, where $N_{t-1} \ll N$ (N the sample size).

For the moment, let us suppose that good upper bounds $c_j(r)$ are known for each r. We discuss later how such bounds can be found. Then, the next step is to evaluate the acceptance probability of a proposed pair (i, z). We find:

Lemma 3.10 *Let the densities $\rho_j(i, z)$ and the distribution $(\tau_j(i))$ be defined as in Lemma 3.8 and Definition 3.5, respectively.*

Then the acceptance probability of the pair (i, z) generated from the distribution $\tau_j(i)\rho_j(i, z)$ is

$$\pi_j(i, z) = \frac{w_{i,j}}{c_j(r)} \; M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) \cdot$$
$$\cdot \exp\left\{ -m \left[\log\left(1 + \left(\frac{z}{c/\widetilde{\sigma}} \right)^2 \right) + \alpha_{k^*} + \beta_{k^*} z + \gamma_{k^*} z^2 \right] \right\}.$$

I_r and B_{k^*} are the subsets containing the index i and the value z, respectively.

Remark 3.7 *The value of $M\left(\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t\right)$ is a by-product of the sampling of Z since $M\left(\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t\right)=\sum_{l=1}^{K}R_{l,ij}M_{l,ij}$, see Remark 3.5. In addition, the resulting acceptance probability has again a very intuitive form. It reflects the two approximations used to implement this method: the approximation of the logarithm of the Pearson type VII density (up to some terms) by a majorant and the grouping of similar indices i to define $(\tau_j(i))$. Therefore, it is again important that the majorant approximates well the logarithm of the Pearson type VII density (up to some terms) and that the upper bounds $c_j(r)$ are chosen with care. In this way, we have a high acceptance probability.*

Proof: The proof is similar to the corresponding ones in the filtering case and in the smoothing recursion implemented with the first method, see Lemmas 2.5 and 3.6. That is we take the formula for the acceptance probability in the rejection method with the auxiliary index and we evaluated it using the proposal densities $\rho_j(i,z)$ and the distribution $(\tau_j(i))$ (see their definitions in Lemma 3.8 and in Definition 3.5). Note that the target density is given by $\widehat{p}\left(z|z^{(j)}_{(t+1):(t+p)},\widetilde{y}_{1:t}\right)$, see (3.10). The acceptance probability is given equivalently to (2.16) by

$$\pi_j(i,z)=\frac{p_{VII}(m,c/\widetilde{\sigma},0)(z)\;\;w_{i,j}\;\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\tau_j(i)\;\rho_j(i,z)\;\;M}$$

with

$$M\geq\sup_{i,z}\frac{p_{VII}(m,c/\widetilde{\sigma},0)(z)\;\;w_{i,j}\;\phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}},1\right)(z)}{\tau_j(i)\;\rho_j(i,z)}.$$

First, we calculate the supremum term and we define M. This can be achieved using again the inequality (3.16) and the definitions of $M(a)$,

$\tau_j(i)$ and $c_j(r)$, see Remark 3.5 and Definition 3.5. Explicitly:

$$\begin{aligned}
&\frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z)\ \ w_{i,j}\ \phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{\tau_j(i)\ \rho_j\left(i, z\right)} \\
&\leq \frac{k_1\ w_{i,j}\ M\left(\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t\right)}{\tau_j(i)} \\
&\leq \sup_r \left(\sup_{i \in I_r} \frac{k_1\ w_{i,j}\ M\left(\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t\right)\ \sum_{r=1}^{N_{t-1}} |I_r|\, c_j(r)}{\sum_{r=1}^{N_{t-1}} c_j(r) \mathbb{I}_{\{i \in I_r\}}}\right) \\
&= k_1 \left(\sum_{r=1}^{N_{t-1}} |I_r|\, c_j(r)\right)\ \sup_r \left[\frac{1}{c_j(r)} \sup_{i \in I_r} \left(w_{i,j} M\left(\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t\right)\right)\right] \\
&\leq k_1 \sum_{r=1}^{N_{t-1}} |I_r|\, c_j(r).
\end{aligned}$$

The last expression does not depend neither on z nor on the filter sample $(z_{(t-p):(t-1)}^{(i)})$. Therefore, we set M equal to it. Then, it follows using the definition of $(\tau_j(i))$ and recalling that i is assumed to be in I_r:

$$\begin{aligned}
\pi_j\left(i, z\right) &= \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z)\ \ w_{i,j}\ \phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{k_1 \left(\sum_{r=1}^{N_{t-1}} |I_r|\, c_j(r)\right)\ \tau_j(i)\ \rho_j\left(i, z\right)} \\
&= \frac{w_{i,j} \sum_{l=1}^{K} R_{l,ij} M_{l,ij}}{c_j(r)}\ \frac{p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z)\ \ \phi\left(\frac{\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t}{\widetilde{\sigma}}, 1\right)(z)}{k_1\ \left(\sum_{l=1}^{K} R_{l,ij} M_{l,ij}\right)\ \rho_j\left(i, z\right)} \\
&= \frac{w_{i,j}\ M\left(\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t\right)}{c_j(r)} \cdot \\
&\quad \cdot \exp\left\{-m\left[\log\left(1+\left(\frac{z}{c/\widetilde{\sigma}}\right)^2\right)+\alpha_{k^*}+\beta_{k^*} z+\gamma_{k^*} z^2\right]\right\}.
\end{aligned}$$

The last equality follows since the second term can be simplified as in Lemma 2.5, equation (2.24). Substitute again σ_V by $\widetilde{\sigma}$ and $\sum_{l=1}^{p} \varphi_l z_{t-l}^{(i)}$

by $\mu_i^{(-)} + \mu_j^{(+)}$, set $\lambda = 1$ and recall that z is supposed to be in B_{k^*}. □

If we look closer to the acceptance probability in Lemma 3.10, we note that it is less than

$$\frac{w_{i,j}\ M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right)}{c_j(r)} \tag{3.17}$$

for all pairs (I, Z). In fact, it follows from the majorant Definition 3.1 that the exponential term in the acceptance probability has always negative exponent and therefore it is at most 1. In addition, the upper bound (3.17) for the acceptance probability depends on the generated index I but not on the sampled value Z. This nice feature suggests us the second main idea to reduce the amount of computations. We can carry out a pretesting (squeezing) on the sampled index I before sampling the variable Z. The squeeze method was proposed by Marsaglia (1977).

In our case, this idea is carried out by setting the pretesting probability of the index I equal to the above bound or an easier upper approximation of it (note that the bound is at most 1 thanks to the definition of the $c_j(r)$'s). If I is not accepted, the pair (I, Z) will also not be accepted independent of the value of the sampled Z since the acceptance probability of the pair (I, Z) cannot exceed the upper bound (3.17). In such a case, we spare the computations to generate a useless Z and we go on sampling an index I until it passes the pretesting. An accepted index I constitutes a reliable index and it is worth sampling Z according to $\rho_j(i, z)$ with $I = i$ (see Lemma 3.8) and then checking the acceptance of the pair (i, z) (see Lemma 3.10). In fact the conditional acceptance probability of the pair (I, Z) given that I has passed the pretesting is

$$\exp\left\{-m\left[\log\left(1 + \left(\frac{z}{c/\widetilde{\sigma}}\right)^2\right) + \alpha_{k^*} + \beta_{k^*} z + \gamma_{k^*} z^2\right]\right\}$$

which is about 1 for a suitable chosen majorant. Moreover, the proposal $\rho_j(i, z)$ is found only for one filter particle $z_{(t-p):(t-1)}^{(i)}$ at a time and not for all values as in the first method. Thus, the construction of $\rho_j(i, z)$ can be carried out using the true $\mu_j^{(+)}$ and not only using the approximate value $\overline{\mu_j^{(+)}}$. This also improves the efficiency of the method and it reduces the amount of computations.

Unfortunately, the evaluation of the upper bound (3.17) requires the computation of $M\left(\mu_i^{(-)}+\mu_j^{(+)}-\widetilde{y}_t\right)$ which is the normalizing constant of $\rho_j\,(i,z)$, see Remark 3.5. We should find a way to compute it or at least to get a good upper bound. A Theorem proved by Ibragimov (1956) help us. Following him we define

Definition 3.6 *A distribution function $F(x)$ is called* unimodal *if there exists one value $x=b$ such that for $x<b$ the function $F(x)$ is convex, while for $x>b$ it is concave.*

A distribution function $F(x)$ is called strong unimodal *if the composition (convolution) of $F(x)$ with any unimodal distribution function is unimodal.*

E is the set of all points x such that $D_dF(x)D_sF(x)\neq 0$ where $F(x)$ is a distribution function, $D_dF(x)$ is the right derivative of $F(x)$ and $D_sF(x)$ is the left one.

A probability density function $f(x)$ is called unimodal *or* strong unimodal *if the corresponding distribution function $F(x)$ is unimodal or strong unimodal, respectively.*

Remark 3.8 *The definition of an unimodal probability density function corresponds to the one for general functions:*
a function $h(x)$ is called unimodal, if for some value b, $h(x)$ is monotonic increasing for $x<b$ and monotonic decreasing for $x>b$.

Theorem 3.1 (Ibragimov) *A proper unimodal distribution function $F(x)$ is strong unimodal if and only if $F(x)$ is continuous and $\psi(x) := \log(F'(x))$ is a concave function on E.*

Proof: See Ibragimov (1956). □

Corollary 3.1 *The normal distribution function is strong unimodal.*

Proof: Let $\Phi_{\mu,\sigma}\,(x)$ denote the $\mathcal{N}(\mu,\sigma^2)$ distribution function. Clearly, $\Phi_{\mu,\sigma}\,(x)$ is unimodal (set $b=\mu$) and continuous. Its first derivative is zero nowhere and $E=\mathbb{R}$. Moreover, the function

$$\psi(x) := \log\left(\frac{d}{dx}\Phi_{\mu,\sigma}\,(x)\right) = -\frac{1}{2}\log(2\pi)-\log\sigma-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

is concave on $\mathbb{R}$. Then, the corollary follows from Theorem 3.1 (Ibragimov). □

Now, we show how this is useful.

Lemma 3.11 *Suppose that the majorant $q(z)$ in Definition 3.3 is bounded and unimodal. Then, the function $\widetilde{M}(a)$ is unimodal.*

Proof: The integral $\int q(z)\,dz$ is not assumed to exist. For example, this integral does not exist for the chosen (logarithmic) majorants (see Definition 3.9) since they do not go to $-\infty$ for x going to $\pm\infty$. Thus, we cannot define and work directly with $q(z) / \int q(z)\,dz$.
The way out is given by the following definition. Let

$$q_n(z) = q(z)\,\mathbb{I}_{\{z\in[-n,n]\}}.$$

Then

$$\int_{\infty}^{\infty} q_n(z)\,dz = \int_{-n}^{n} q(z)\,dz \leq \; 2n \;\sup_z q(z) < \infty.$$

Therefore, we can define the density

$$f_n(z) = \frac{q_n(z)}{\int q_n(z)\,dz}.$$

This density $f_n(z)$ is unimodal. Thus, the convolution $(F_n * \Phi)(a)$ is unimodal from Corollary 3.1 where $F_n(z)$ denotes the distribution function corresponding to $f_n(z)$ and $\Phi(z)$ denotes the standard normal distribution function. Since

$$\frac{d}{da}(F_n * \Phi)(a) = \frac{d}{da}\int \Phi(a-z)\,dF_n(z) = \int \phi_1(a-z)\,\frac{q_n(z)}{\int q_n(z)\,dz}dz,$$

the function

$$\widetilde{M}_n(a) := \frac{1}{k_1}\int \phi_1(a-z)\,q_n(z)\,dz$$

is unimodal for all n. Finally, from $q_n(z) \leq q(z) \leq \sup_z q(z) < \infty$, it follows using Lebesgue's theorem that

$$\lim_{n\to\infty} \widetilde{M}_n(a) = \frac{1}{k_1}\int \phi_1(a-z)\,q(z)\,dz = \widetilde{M}(a), \quad \forall a$$

and the lemma is proved. □

Clearly, $M(a) = \widetilde{M}(a/\widetilde{\sigma})$ is also unimodal and we denote the maximum of $M(a)$ by a_{max}. We can benefit from this unimodal feature to compute an upper bound for $M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right)$. In fact, we can evaluate $M(a)$ on a grid $a_1, \ldots a_{N_g}$ at the beginning of the smoothing algorithm. Then, we approximate $M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right)$ by taking the smallest value $M(a_l)$ greater than $M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right)$, a_l a grid point, or $M(a_{max})$. As a consequence of the unimodal feature, only a_{max} or two grid points come into question: the largest a_l not greater than $\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t$ or the smallest a_l not less than $\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t$. If we would know the location of the maximum a_{max}, then we could choose the right point between them. But unfortunately, it is not easy to find the position of a_{max} for a general majorant $q(z)$. We should refine this idea and assume additionally that the majorant $q(z)$ is symmetric about zero. Then, the same is true for $M(a)$ (see Remark (3.5)) and therefore a_{max} is zero.

Explicitly, we adopt the following technique to get an upper bound for $M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right)$. First:

Definition 3.7 *Let $a_1, \ldots, a_{N_g}$ be a grid containing zero and symmetric about it. Then, we compute $M(a_l)$ for all grid points a_l not greater than zero by setting $\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t := a_l$ in the definitions of $R_{k,ij}$ and $M_{k,ij}$ (see Lemma 3.8) to get*

$$M(a_l) = M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) = \sum_{k=1}^{K} R_{k,ij} M_{k,ij}.$$

For positive grid points a_l, it follows from the symmetries of both the function $M(a)$ and the grid sequence that

$$M(a_l) = M(-a_l) = M\left(a_{N_g+1-l}\right).$$

The next step is to compute $\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t$. It follows:

- If $\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t \leq 0$, then $M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) \leq M(a_{l*})$ with
$$a_{l*-1} < \mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t \leq a_{l*}. \tag{3.18}$$

- If $\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t > 0$, then $M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) \leq M\left(a_{l^*}\right)$ with
$$a_{l^*} \leq \mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t < a_{l^*+1}. \tag{3.19}$$

Moreover, the grid idea permits to answer the open question of finding the $c_j(r)$'s for all $r = 1, \ldots, N_{t-1}$. They have been defined by

$$c_j(r) \geq \sup_{i \in I_r} \left(w_{i,j}\ M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right)\right),$$

see Definition 3.5, and recall that $\mu_j^{(+)}$ is fixed. It is important to find a good approximation of the supremum to define $c_j(r)$ since the quotient

$$\frac{w_{i,j}\ M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right)}{c_j(r)}$$

is the acceptance probability of the index I in the pretesting. Clearly, it is desirable to have a high acceptance probability. But on the other hand, the computation of the $c_j(r)$'s has to be easy to carry out. Then, the most intuitive idea is to set

$$c_j(r) \geq \sup_{i \in I_r} w_{i,j} \ \cdot \ \sup_{i \in I_r} M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right).$$

How can we find the two suprema? As before, an upper bound for the latter one can be computed thanks to the grid.

Definition 3.8 *Let*

$$\mu(r)_{min} := \min_{i \in I_r}\left(\mu_i^{(-)}\right) + \mu_j^{(+)} - \widetilde{y}_t,$$
$$\mu(r)_{max} := \max_{i \in I_r}\left(\mu_i^{(-)}\right) + \mu_j^{(+)} - \widetilde{y}_t.$$

First, we note that $\mu(r)_{min} \leq \mu(r)_{max}$. Then:

- If $\mu(r)_{max} \leq 0$, then
$$\sup_{i \in I_r} M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) = M\left(\mu(r)_{max}\right) \leq M\left(a_{l^*}\right)$$
with
$$a_{l^*-1} < \mu(r)_{max} \leq a_{l^*}. \tag{3.20}$$

- If $\mu(r)_{min} \geq 0$, then

$$\sup_{i \in I_r} M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) = M\left(\mu(r)_{min}\right) \leq M\left(a_{l^*}\right)$$

  with

$$a_{l^*} \leq \mu(r)_{min} < a_{l^*+1}. \tag{3.21}$$

- Otherwise, we have $\mu(r)_{min} < 0 < \mu(r)_{max}$ and then

$$\sup_{i \in I_r} M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right) = M\left(a_{l^*}\right)$$

  with

$$a_{l^*} = 0 \tag{3.22}$$

  (recall that 0 is a grid point).

For the first supremum, it follows using (3.9) that

$$\begin{aligned} \sup_{i \in I_r} w_{i,j} &= \sup_{i \in I_r} \left[k \ \exp\left(-\frac{1}{2}\left\|Rv_j^{(+)} - Rv_i^{(-)}\right\|_2^2\right)\right] \\ &= k \exp\left(-\frac{1}{2}\left\|Rv_j^{(+)} - Rv_{i^*}^{(-)}\right\|_2^2\right) =: w_{i^*,j} \end{aligned}$$

where $Rv_{i^*}^{(-)}$ denotes the nearest neighbour of the given $Rv_j^{(+)}$ in the subset $\{Rv_i^{(-)}, i \in I_r\}$. The intuitive procedure to find $Rv_{i^*}^{(-)}$ is to compute the distances between $Rv_j^{(+)}$ and all $Rv_i^{(-)}$'s, $i \in I_r$, and return the $Rv_{i^*}^{(-)}$ with the minimal distance. Since the procedure is applied to all subsets I_r and all vectors $Rv_j^{(+)}$ in each smoothing step, the resulting algorithm will have a complexity of order TN^2. We saw that the first method to implement the smoothing recursion had the same complexity and the aim of the second idea is to reduce it. This can be achieved with the method proposed by Friedman et al. (1977). Briefly, they organized the subset $\{Rv_i^{(-)}, i \in I_r\}$ in a tree structure and then they searched the nearest neighbour of the query vector $Rv_j^{(+)}$ in this tree. The computations required to construct the tree and to perform the search are proportional to $p\,|I_r|\log(|I_r|)$ and $\log(|I_r|)$, respectively (p is the order of the AR process used as state equation (2.6)). Particularly, the search of the neighbour is extremely fast once the data set is organized. Thus, we adopt the latter nearest neighbour algorithm since

it has a lower complexity compared to the intuitive one and therefore it is considerably faster. Explicitly, we use the C++ implementation written by S. Arya and D. Mount and described in Arya et al. (1998). It supports any Minkowski norm for defining the distance and, especially, the Euclidean one. In addition, it is convenient for us to find the nearest neighbours of all $Rv_j^{(+)}$'s in the subset $\{Rv_i^{(-)}, i \in I_r\}$ at the same time and then iterate for all subsets.
Summarizing, the $c_j(r)$'s are defined for each $r = 1, \dots, N_{t-1}$ by

$$c_j(r) = w_{i^*,j} \cdot M\left(a_{l^*}\right) \tag{3.23}$$

where $w_{i^*,j}$ and a_{l^*} are defined as above. We remarked before that each $c_j(r)$ should be preferably close to $\left\{w_{i,j} M\left(\mu_i^{(-)} + \mu_j^{(+)} - \widetilde{y}_t\right), i \in I_r\right\}$ to have a high acceptance probability in the pretesting for I. This can be reached by choosing a fine grid $a_1, \dots a_{N_g}$ and a small width for the subsets S_r. Of course, this can result in an increment of the computations and thus a good trade-off should be found in each situation.

Finally, we examine the construction of the (logarithmic) majorant needed to find the proposals $\rho_j\left(i, z\right)$, see Definition 3.1. In the previous paragraphes, we have assumed that the majorant $q(z)$ of the density $p_{VII}\left(m, c/\widetilde{\sigma}, 0\right)(z)$ is symmetric about zero, bounded and unimodal. We needed these features to carry out the computations. Therefore, we should take care that the constructed (logarithmic) majorant fulfils these features.
Actually, as before, we distinguish three cases. Note that the lower and the upper majorants should be defined differently.

Definition 3.9 *Let the $\mu_i^{(-)}$'s, the $\mu_j^{(+)}$'s and $\widetilde{\sigma}$ be as in (3.5), (3.6) and Lemma 3.1. In addition, let δ be a strictly positive real number. Define for a fixed $\mu_j^{(+)}$:*

$$\mu = \frac{med_i\left(\mu_i^{(-)}\right) + \mu_j^{(+)} - \widetilde{y}_t}{\widetilde{\sigma}},$$

$$l_{(-)} = g(\mu - \delta) = -\log\left[1 + \left(\frac{\mu - \delta}{c/\widetilde{\sigma}}\right)^2\right],$$

$$l_{(+)} = g(\mu + \delta) = -\log\left[1 + \left(\frac{\mu + \delta}{c/\widetilde{\sigma}}\right)^2\right].$$

Then, the parameters to define the default majorant are given in Table 3.4, the parameters for the lower majorant in Table 3.5 and the parameters for the upper majorant in Table 3.6. Note that lower and upper majorants can be defined only if $\mu < -\sqrt{3}\ c/\widetilde{\sigma} - \delta$ *and* $\mu > \sqrt{3}\ c/\widetilde{\sigma} + \delta$*, respectively.*

k	d_{k-1}	d_k	α_k	β_k	γ_k
1	$-\infty$	$-5\ c/\widetilde{\sigma}$	$-\log(26)$	0	0
2	$-5\ c/\widetilde{\sigma}$	$-\sqrt{3}\ c/\widetilde{\sigma}$	$\frac{\sqrt{3}\log(6.5)}{5-\sqrt{3}} - \log(4)$	$\frac{\log(6.5)}{\left(5-\sqrt{3}\right)\ c/\widetilde{\sigma}}$	0
3	$-\sqrt{3}\ c/\widetilde{\sigma}$	$-c/\widetilde{\sigma}$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\widetilde{\sigma}}$	0
4	$-c/\widetilde{\sigma}$	$c/\widetilde{\sigma}$	0	0	$-\frac{\log(2)}{(c/\widetilde{\sigma})^2}$
5	$c/\widetilde{\sigma}$	$\sqrt{3}\ c/\widetilde{\sigma}$	$\frac{\log(4)-\sqrt{3}\log(2)}{\sqrt{3}-1}$	$-\frac{\log(2)}{\left(\sqrt{3}-1\right)\ c/\widetilde{\sigma}}$	0
6	$\sqrt{3}\ c/\widetilde{\sigma}$	$5\ c/\widetilde{\sigma}$	$\frac{\sqrt{3}\log(6.5)}{5-\sqrt{3}} - \log(4)$	$-\frac{\log(6.5)}{\left(5-\sqrt{3}\right)\ c/\widetilde{\sigma}}$	0
7	$5\ c/\widetilde{\sigma}$	∞	$-\log(26)$	0	0

Table 3.4: *Parameters to define the default majorant in the smoothing recursion (second method).*

k	d_{k-1}	d_k	α_k	β_k	γ_k
1	$-\infty$	$\mu-\delta$	$l_{(-)}$	0	0
2	$\mu-\delta$	$\mu+\delta$	$l_{(-)} - \beta_2\ (\mu-\delta)$	$\frac{l_{(+)}-l_{(-)}}{2\delta}$	0
3	$\mu+\delta$	$-\sqrt{3}\ c/\widetilde{\sigma}$	$-\log(4) + \beta_3\ \sqrt{3}\ c/\widetilde{\sigma}$	$\frac{\log(4)+l_{(+)}}{\mu+\delta+\sqrt{3}\ c/\widetilde{\sigma}}$	0
4	$-\sqrt{3}\ c/\widetilde{\sigma}$	$\sqrt{3}\ c/\widetilde{\sigma}$	0	0	$-\frac{\log(4)}{3(c/\widetilde{\sigma})^2}$
5	$\sqrt{3}\ c/\widetilde{\sigma}$	$-\mu-\delta$	$-\log(4) - \beta_5\ \sqrt{3}\ c/\widetilde{\sigma}$	$-\frac{\log(4)+l_{(+)}}{\mu+\delta+\sqrt{3}\ c/\widetilde{\sigma}}$	0
6	$-\mu-\delta$	$-\mu+\delta$	$l_{(-)} + \beta_6\ (\mu-\delta)$	$-\frac{l_{(+)}-l_{(-)}}{2\delta}$	0
7	$-\mu+\delta$	∞	$l_{(-)}$	0	0

Table 3.5: *Parameters to define the lower majorant in the smoothing recursion (second method). We should have* $\mu < -\sqrt{3}\ c/\widetilde{\sigma} - \delta$.

k	d_{k-1}	d_k	α_k	β_k	γ_k
1	$-\infty$	$-\mu-\delta$	$l_{(+)}$	0	0
2	$-\mu-\delta$	$-\mu+\delta$	$l_{(+)}+\beta_2\ (\mu+\delta)$	$\frac{l_{(-)}-l_{(+)}}{2\delta}$	0
3	$-\mu+\delta$	$-\sqrt{3}\ c/\widetilde{\sigma}$	$-\log(4)+\beta_3\ \sqrt{3}\ c/\widetilde{\sigma}$	$\frac{l_{(-)}+\log(4)}{\sqrt{3}\ c/\widetilde{\sigma}-\mu+\delta}$	0
4	$-\sqrt{3}\ c/\widetilde{\sigma}$	$\sqrt{3}\ c/\widetilde{\sigma}$	0	0	$-\frac{\log(4)}{3(c/\widetilde{\sigma})^2}$
5	$\sqrt{3}\ c/\widetilde{\sigma}$	$\mu-\delta$	$-\log(4)-\beta_5\ \sqrt{3}\ c/\widetilde{\sigma}$	$-\frac{l_{(-)}+\log(4)}{\sqrt{3}\ c/\widetilde{\sigma}-\mu+\delta}$	0
6	$\mu-\delta$	$\mu+\delta$	$l_{(+)}-\beta_6\ (\mu+\delta)$	$-\frac{l_{(-)}-l_{(+)}}{2\delta}$	0
7	$\mu+\delta$	∞	$l_{(+)}$	0	0

Table 3.6: *Parameters to define the upper majorant in the smoothing recursion (second method). We should have* $\mu > \sqrt{3}\ c/\widetilde{\sigma}+\delta$

We select:

- *the default majorant if* $-\sqrt{3}\ c/\widetilde{\sigma}-\delta \leq \mu \leq \sqrt{3}\ c/\widetilde{\sigma}+\delta,$
- *the lower majorant if* $-\sqrt{3}\ c/\widetilde{\sigma}-\delta > \mu,$
- *the upper majorant if* $\mu > \sqrt{3}\ c/\widetilde{\sigma}+\delta.$

Remark 3.9 *Most of the remarks done in the filtering case are also valid here, see Remark 2.8.*
New is the feature that lower and upper majorants are symmetric about zero, too. Consequently, we have to simplify somewhat their approximations around zero to retain the same number of subsets as before (7 subsets). Note that the upper majorant can be obtained from the lower one by setting $\mu_{upp} = -\mu_{low}$.
In addition, we need to evaluate $M(a)$ *on the grid points* $a_1, \ldots, a_{N_g}$ *to perform the smoothing recursion. But the computation of* $M(a_l)$ *depends on the chosen majorant, see Definition 3.7. Thus,* $M(a_l)$ *can be computed for the default majorant at the beginning of the smoothing algorithm. But, every time that the lower or upper majorant is chosen, we would have to compute* $M(a_l)$ *brand-new. We can avoid partially this disadvantage using the same idea as in the first smoothing method. We sort the values* $\mu_j^{(+)}$ *and we use the same lower or upper majorant and the corresponding* $M(a_l)$*'s possibly for several* $\mu_j^{(+)}$*'s. When the lower or upper majorant is no more suitable, we compute a better one and also the corresponding* $M(a_l)$*'s. The algorithm 3.3 explains the procedure in detail.*

Lemma 3.12 *The three majorants in Definition 3.9 satisfy Definition 3.1. I.e., they fulfil the inequality*

$$g(z) := -\log\left[1+\left(\frac{z}{c/\widetilde{\sigma}}\right)^2\right] \leq \sum_{k=1}^{K}\left(\alpha_k+\beta_k z+\gamma_k z^2\right)\mathbb{I}_{\{z\in B_k\}},\ \forall z\in\mathbb{R}.$$

The three majorants are continuous functions in z. Moreover, if $q\,(z)$ is constructed from one of them, then $q\,(z)$ is symmetric about zero, bounded and unimodal.

Proof: The proofs of both the inequality and the continuity are similar to the ones in the filtering case, see Lemma 2.6. Substitute σ_V by $\widetilde{\sigma}$ and, for the lower and upper majorants, prove the inequality on the interval $[-\sqrt{3}, \sqrt{3}]$ similarly as before.
In addition, let $h(z)$ denote the chosen majorant. We define $q\,(z)$ by

$$q\,(z) = \frac{\Gamma(m)\ \widetilde{\sigma}}{\sqrt{\pi}\ c\,\Gamma(m-0.5)}\ \exp\left(mh(z)\right).$$

Then, $q\,(z)$ is symmetric about zero, bounded and unimodal. □

Finally, we put together all previous results and we summarize the particle smoothing recursion at time t with known $\widetilde{y}_t$.

Algorithm 3.3 *Particle smoothing recursion at time t with $\widetilde{y}_t$ available.*
Assumptions:

- *The fully defined default majorant and the partially defined lower and upper majorants have already been computed and δ is known, see Definition 3.9.*
- *The grid $a_1, \ldots, a_{N_g}$ has already been defined: it contains zero and it is symmetric about it. In addition, all $M\,(a_l)$'s have been evaluated using the default majorant, see Definition 3.7.*
- *The upper triangular matrix R has already been found as explained in Remark 3.1.*
- *The width of the subsets that divide the range $\left[\min\left(\mu_i^{(-)}\right), \max\left(\mu_i^{(-)}\right)\right]$ and the maximal allowed number of rejections are known.*

- *The filter sample* $(z^{(i)}_{(t-p):(t-1)})$ *and the smoothing particles* $(z^{(j)}_{(t+1):(t+p)})$, $i,j = 1,\dots,N$, *are known from the filtering recursion and the previous smoothing step at time* $t+1$, *respectively.*

Preliminaries:

1. *Compute the* $\mu^{(-)}_i$*'s and the* $\mu^{(+)}_j$*'s for* $i,j=1,\dots,N$ *as described in (3.5), (3.6) and Lemma 3.1.*
   Compute the $Rv^{(-)}_i$*'s and the* $Rv^{(+)}_j$*'s for* $i,j=1,\dots,N$ *with* R *as above and* $v^{(-)}_i$*'s,* $v^{(+)}_j$*'s given as in (3.7), (3.8) and Lemma 3.1.*
   Sort the $\mu^{(-)}_i$*'s according to size and apply the same permutation to the* $Rv^{(-)}_i$*'s. Do the same for the* $\mu^{(+)}_j$*'s and the* $Rv^{(+)}_j$*'s.*
2. *Divide the range* $\left[\min\left(\mu^{(-)}_i\right), \max\left(\mu^{(-)}_i\right)\right]$ *in subsets* $S_1, \dots, S_{N_{t-1}}$ *with the given width and compute all* $I_r := \left\{i \middle| \mu^{(-)}_i \in S_r\right\}$, $r = 1,\dots,N_{t-1}$.
   Compute the nearest neighbour of each $Rv^{(+)}_j$ *in all subsets* $\{Rv^{(-)}_i, i \in I_r\}$, $r=1,\dots,N_{t-1}$.

Begin the construction of the particles $(z^{(j)}_t)$, $j=1,\dots,N$. *Set* $j=1$ *and initialize the rejection counter* rej-counter *to 0.*

3. *Choose the majorant to generate* $z^{(j)}_t$.
   For this purpose, define
   $$\mu := \frac{med_i\left(\mu^{(-)}_i\right) + \mu^{(+)}_j - \widetilde{y}_t}{\widetilde{\sigma}}$$
   and select
   - *the default majorant if* $-\sqrt{3}\ c/\widetilde{\sigma} - \delta \le \mu \le \sqrt{3}\ c/\widetilde{\sigma} + \delta$,
   - *the lower majorant if* $-\sqrt{3}\ c/\widetilde{\sigma} - \delta > \mu$,
   - *the upper majorant if* $\mu > \sqrt{3}\ c/\widetilde{\sigma} + \delta$.

   In the cases where the lower or the upper majorant is chosen:
   - *If it has been chosen in this smoothing step for the first time or* rej-counter *is greater than the maximal allowed number of rejections, then complete the definition of the lower or the*

upper majorant computing the components which depend on μ and δ. In addition, evaluate all $M(a_l)$'s, $l = 1, \ldots, N_g$, with this new majorant (see Definition 3.7).

- *Otherwise, use the previous defined lower or upper majorant and the corresponding $M(a_l)$'s.*

Set rej-counter *to* 0.

4. *Compute the $c_j(r)$'s, $r = 1, \ldots, N_{t-1}$, and the distribution $(\tau_j(i))$.*

   Explicitly, from (3.23):

$$c_j(r) := w_{i^*,j} \cdot M(a_{l^*}) \text{ with } w_{i^*,j} = k \exp\left(-\frac{1}{2}\left\|Rv_j^{(+)} - Rv_{i^*}^{(-)}\right\|_2^2\right).$$

   $Rv_{i^*}^{(-)}$ is the nearest neighbour of $Rv_j^{(+)}$ in the subset $\{Rv_i^{(-)}, i \in I_r\}$ and a_{l^*} is as in (3.20), (3.21) or (3.22). The distribution $(\tau_j(i))$ is found as in Definition 3.5. In addition, compute also the partial sums of the latter distribution.

5. *Sample an index $I^{(j)}$ according to $(\tau_j(i))$ until it passes the pretesting.*

   To this end, sample $I^{(j)}$ with the inversion method and evaluate the pretesting probability $\pi_j\left(i^{(j)}\right)$:

$$\pi_j\left(i^{(j)}\right) := \frac{w_{i^{(j)},j}\ M(a_{l^*})}{c_j(r)}$$

   where I_r is the subset containing $i^{(j)}$ and $M(a_{l^*})$ is as in (3.18) or (3.19). Then, generate U uniform on $[0, 1]$.

   - *If $U > \pi_j\left(i^{(j)}\right)$, then the pair $(I^{(j)}, Z^{(j)})$ will be certainly rejected ($Z^{(j)}$ is generated according to $\rho_j\left(i^{(j)}, z\right)$). Increase* rej-counter *by* 1 *and:*
     - *If the default majorant has been chosen: repeat the sampling of $I^{(j)}$.*
     - *If the lower or the upper majorant has been chosen:*
       if rej-counter *is not greater than the maximal permitted number of rejections or the lower or the upper majorant has already been computed for this j: repeat the sampling of $I^{(j)}$.*
       Otherwise, go back to step 3 (improve the majorant).
   - *Else, $I^{(j)}$ passes the pretesting. Go to step 6.*

6. *Sample $Z^{(j)}$ according to $\rho_j\left(i^{(j)}, z\right)$ where $I^{(j)} = i^{(j)}$ has passed the pretesting.*

   For this purpose, construct $\rho_j\left(i^{(j)}, z\right)$ as explained in Lemma 3.8 (note that this is done only for $I^{(j)} = i^{(j)}$ and that the partial sums of the $RM_{k,i^{(j)}j}$'s over k are also needed). The sampling of $Z^{(j)}$ is carried out as explained in Remark 2.6.

7. *Check the acceptance of the proposed pair $\left(i^{(j)}, z^{(j)}\right)$.*

   To this end, compute the acceptance probability $\pi_j\left(i^{(j)}, z^{(j)}\right)$ according to Lemma 3.10. Then, consider the same U for which $I^{(j)}$ has passed the pretesting.

   - *If $U \leq \pi_j\left(i^{(j)}, z^{(j)}\right)$, then accept the pair $\left(i^{(j)}, z^{(j)}\right)$. Return the particle $z_t^{(j)}$ defined by*

     $$z_t^{(j)} = \widetilde{y}_t + \widetilde{\sigma}\, z^{(j)}.$$

     Set $j = j + 1$.
     If $j \leq N$, return to step 3. Otherwise stop: all particles $(z_t^{(j)})$ have been computed.
   - *Else, the pair $\left(i^{(j)}, z^{(j)}\right)$ is not accepted. Increment* rej-counter *by 1 and:*
     - *If the default majorant has been chosen: go back to step 5.*
     - *If the lower or the upper majorant has been chosen:*
       if rej-counter *is not greater than the maximal permitted number of rejections or the lower or the upper majorant has already been computed for this j: go back to step 5. Otherwise, go back to step 3 (improve the majorant).*

3.4.2 Recursion with missing $\widetilde{y}_t$

The second case to discuss is the smoothing recursion with missing $\widetilde{y}_t$. As before, this is a special case of the one with $\widetilde{y}_t$ available, and therefore the recursion simplifies considerably.

The discussion in Subsection 3.3.2 remains valid. For each $z^{(j)}_{(t+1):(t+p)}$, the aim is to generate a particle $z_t^{(j)}$ from the density

$\widehat{p}\left(z_t|z^{(j)}_{(t+1):(t+p)},\widetilde{y}_{1:t}\right)$ defined by

$$\widehat{p}\left(z_t|z^{(j)}_{(t+1):(t+p)},\widetilde{y}_{1:t}\right)=\sum_{i=1}^{N}\frac{w_{i,j}}{\sum_{i=1}^{N}w_{i,j}}\;\phi\left(\mu_i^{(-)}+\mu_j^{(+)},\widetilde{\sigma}\right)(z_t)\,.$$

Note that we have not dropped $\widetilde{y}_t$ from the previous smoothing density notation, although it is missing. We have set the "value" of $\widetilde{y}_t$ to NA (*not available*). In this way, the formula has a better readability.

For each $z^{(j)}_{(t+1):(t+p)}$, the above mentioned sampling can be achieved again by first generating the index I from the weight distribution $w_{i,j}/\sum_{i=1}^{N}w_{i,j}$ and then the variable Z_t according to the normal density $\phi\left(\mu_i^{(-)}+\mu_j^{(+)},\widetilde{\sigma}\right)(z_t)$ with $I=i$. We have already remarked that both the weights and the normal densities depend on the filter and the smoothing samples. These dependences do not cause any difficulties to the sampling of Z_t. On the other hand, the dependence on the smoothing sample makes problematic the sampling according to the weight distribution, see the first smoothing method (Subsection 3.3.2). But, this latter sampling can be improved by applying the same idea as in the case with $\widetilde{y}_t$ available. That is we compute the weight distribution for each smoothing particle $z^{(j)}_{(t+1):(t+p)}$ but we group similar filter particles. In this way, the weight distribution is reliable and fast to compute at the same time. The grouping is achieved again by partitioning the range $\left[\min\left(\mu_i^{(-)}\right),\max\left(\mu_i^{(-)}\right)\right]$, although now the weights $w_{i,j}/\sum_{i=1}^{N}w_{i,j}$ do not depend directly on the values $\mu_i^{(-)}$, see (3.9). Another possibility would be to group similar vectors $Rv_i^{(-)}$. But since they are p-dimensional, we risk to meet again the problem of the curse of dimensionality as in the first smoothing method.

Explicitly, the range $\left[\min\left(\mu_i^{(-)}\right),\max\left(\mu_i^{(-)}\right)\right]$ is divided as in Definition 3.4. Then, we simplify Definition 3.5 to find the distribution $(\tau_j(i))$.

Definition 3.10 *The used distribution $(\tau_j(i))$ is defined by*

$$\tau_j(i)=\tau_j(i(r))=\frac{c_j(r)}{\sum_{r=1}^{N_{t-1}}|I_r|\,c_j(r)}$$

with $i\in I_r, r\in\{1,\ldots,N_{t-1}\}$. $c_j(r)$ is defined by

$$c_j(r)=\sup_{i\in I_r}w_{i,j}=k\;\exp\left(-\frac{1}{2}\left\|Rv_j^{(+)}-Rv_{i*}^{(-)}\right\|_2^2\right).$$

$Rv_{i*}^{(-)}$ is the nearest neighbour of $Rv_j^{(+)}$ in the subset $\{Rv_i^{(-)}, i \in I_r\}$.

The sampling of the index I according to the weight distribution can be carried out by the rejection method, see Subsection 2.2.1. I is proposed according to the distribution $(\tau_j(i))$ and it is accepted with probability

$$\pi_j(i) = \frac{\frac{w_{i,j}}{c_j(r)}}{\sup_i \frac{w_{i,j}}{c_j(r)}} = \frac{\frac{w_{i,j}}{c_j(r)}}{\sup_r \sup_{i \in I_r} \frac{w_{i,j}}{c_j(r)}} = \frac{w_{i,j}}{c_j(r)}.$$

Remark 3.10 *If $c_j(r) \approx w_{i,j}$, then the acceptance probability is high. This can be reached by dividing the range $\left[\min\left(\mu_i^{(-)}\right), \max\left(\mu_i^{(-)}\right)\right]$ in small subsets. Of course, more computations are needed.*

Consequently, the particle recursion is organized as follows.

Algorithm 3.4 *Particle smoothing recursion at time t with missing $\widetilde{y}_t$.*
Assumptions:

- *The upper triangular matrix R has already been found as explained in Remark 3.1.*
- *The width of the subsets that divide the range $\left[\min\left(\mu_i^{(-)}\right), \max\left(\mu_i^{(-)}\right)\right]$ is known.*
- *The filter sample $(z_{(t-p):(t-1)}^{(i)})$ and the smoothing particles $(z_{(t+1):(t+p)}^{(j)})$, $i, j = 1, \ldots, N$, are known from the filtering recursion and the previous smoothing step at time $t+1$, respectively.*

Preliminaries:

1. *Compute the $\mu_i^{(-)}$'s and the $\mu_j^{(+)}$'s for $i, j = 1, \ldots, N$ as described in (3.5), (3.6) and Lemma 3.1.*
   Compute the $Rv_i^{(-)}$'s and the $Rv_j^{(+)}$'s for $i, j = 1, \ldots, N$ with R as above and $v_i^{(-)}$'s, $v_j^{(+)}$'s given as in (3.7), (3.8) and Lemma 3.1.
   Sort the $\mu_i^{(-)}$'s according to size and apply the same permutation to the $Rv_i^{(-)}$'s.
2. *Divide the range $\left[\min\left(\mu_i^{(-)}\right), \max\left(\mu_i^{(-)}\right)\right]$ in subsets $S_1, \ldots, S_{N_{t-1}}$ with the given width and compute all*

$I_r := \left\{ i | \mu_i^{(-)} \in S_r \right\}$, $r = 1, \ldots, N_{t-1}$.
Compute the nearest neighbour of each $Rv_j^{(+)}$ *in all subsets* $\{Rv_i^{(-)}, i \in I_r\}$, $r = 1, \ldots, N_{t-1}$.

Begin the construction of the particles $(z_t^{(j)})$, $j = 1, \ldots, N$.

3. *For j from 1 to N do:*

   (a) *Compute the* $c_j(r)$*'s,* $r = 1, \ldots, N_{t-1}$, *and find the distribution* $(\tau_j(i))$.
   Explicitly, from Definition 3.10:

$$c_j(r) := k \, \exp\left(-\frac{1}{2} \left\| Rv_j^{(+)} - Rv_{i*}^{(-)} \right\|_2^2 \right)$$

   with $Rv_{i*}^{(-)}$ *the nearest neighbour of* $Rv_j^{(+)}$ *in the subset* $\{Rv_i^{(-)}, i \in I_r\}$. *Then, find the distribution* $(\tau_j(i))$ *and its partial sums.*

   (b) *Sample the index* $I^{(j)}$ *according to the weight distribution* $w_{i,j} / \sum_{i=1}^{N} w_{i,j}$ *using the rejection method.*
   To this end, $I^{(j)}$ *is proposed according to the distribution* $(\tau_j(i))$ *using the inversion method. Then, evaluate its acceptance probability* $\pi_j\left(i^{(j)}\right)$ *in the rejection method:*

$$\pi_j\left(i^{(j)}\right) = \frac{w_{i^{(j)},j}}{c_j(r)}$$

   with I_r *the subset containing the index* $i^{(j)}$. *Generate* U *uniform on* $[0, 1]$.

   - *If* $U \leq \pi_j\left(i^{(j)}\right)$, *accept* $I^{(j)}$ *and go to the next step.*
   - *Else, repeat the sampling of* $I^{(j)}$.

   (c) *Sample the variable* $Z_t^{(j)}$ *according to the normal density* $\phi\left(\mu_{i^{(j)}}^{(-)} + \mu_j^{(+)}, \widetilde{\sigma}\right)(z_t)$ *with* $I^{(j)} = i^{(j)}$ *and return* $z_t^{(j)}$.

Chapter 4

Parameter estimation

In the previous chapters, we assumed that the whole set of parameters, say θ, in the considered model (2.1) and (2.2) is known and we discussed the inference about the states $Z_{1:T}$ based on the observations $Y_{1:T}$. From a statistical point of view, the most interesting problem is the inference about the unknown parameters in the model. These can be equal to all parameters or just be a subset of them (recall that the parameters consist of both the hyperparameters of the function f and the nuisance parameters determining the distributions of Z_t and ε_t). We propose to estimate the unknown parameters using the maximum likelihood method. This should result in robust estimates since the observation error distribution in (2.2) is assumed to be heavy-tailed. In addition, the maximum likelihood approach permits to find approximate confidence intervals by the usual likelihood techniques. The difficulty is that the likelihood function cannot be evaluated in closed form. Thus, it should be approximated using Monte Carlo methods. For this reason, we considered the particle filtering and smoothing recursions in Chapters 2 and 3. In fact, fast and reliable algorithms for these recursions are a prerequisite for computing maximum likelihood estimators.

We discuss in this chapter two maximum likelihood methods to estimate the unknown parameters. We call them the smoothing method and MCEM. Both approaches have already been introduced by Hürzeler (1998), see Chapter 7. We will generalize these methods to cover also the case with missing values in the observations (Y_t) and/or in the external regressors $(X_{t,1}, \ldots, X_{t,m})$. In addition, the methods are protected

against numeric overflow.
The key idea of both approaches is to approximate the exact likelihood function using smoothing particles computed for a given θ_0. (Recall that θ_0 consists of both the unknown parameters and the given ones.) Then, the approximate likelihood functions are maximized with respect to the unknown parameters. Since the approximations are reliable only around the used θ_0, the procedures are iterated until the parameter estimates are "stable" according to a chosen criterion.
We already discuss some features of this idea. A disadvantage is that the filtering and smoothing recursions should be computed in each iteration. For this reason, we made the effort to develop fast and reliable algorithms in the previous chapters. On the other hand, the two methods have a complexity of order TN. Therefore, the iterations are fast to compute, especially for small sample sizes N. Moreover, the iterative estimates have quite small fluctuations around the true maximum likelihood value $\widehat{\theta}_{ML}$ already with small values of N. Unfortunately, their convergence towards $\widehat{\theta}_{ML}$ may be slow. This last behaviour can be improved easily by developing an algorithm to find good starting estimates for the unknown parameters. In this way, the number of iterations needed decreases and the estimation algorithms are faster. The random fluctuations of the iterative estimates around $\widehat{\theta}_{ML}$ can be reduced by increasing the sample size N. It is reasonable to start the recursions with a small value of N, and then increment it during the iterations. Finally, we can use an ad hoc idea as stopping criterion. The estimate differences $\widehat{\theta}_k - \widehat{\theta}_{k-1}$ are plotted against k and the index k from when the differences fluctuate around zero is found by eye.

Hürzeler (1998) also proposed a maximum likelihood method based on filter particles, see again Chapter 7. An advantage is of course that the smoothing recursion is not carried out. In addition, the convergence is achieved with few iterations. But unfortunately, the iterative estimates have big fluctuations around $\widehat{\theta}_{ML}$. Their amplitude could be reduced by enlarging the sample size N. But this is not a good idea since the algorithm has the disadvantage to have a complexity of order TN^2. Thus, it would become very slow.
For this reason, we give up to implement the filtering method. With a good algorithm to compute the starting estimates, the smoothing and the MCEM methods are more reliable and faster.

A definition and a remark before we begin the detailed discussion. In this chapter, we consider explicitly the times t for which both the observation Y_t and the external regressors $(X_{t,1}, \ldots, X_{t,m})$ are available.

(Recall that in Chapters 2 and 3, we gave the value NA to missing observations or missing external regressors. In this way, formulae had a better readability.)

Definition 4.1 *Let the time index subset I_{av} be defined by*

$$I_{av} = \{t \mid Y_t \textit{ and } (X_{t,1}, \ldots, X_{t,m}) \textit{ are available}\} \subseteq \{1, \ldots, T\}.$$

Moreover, since we are looking for the log-likelihood function, we work with (Y_t) and not with $(\widetilde{Y}_t)$. In fact, $(\widetilde{Y}_t)$ were defined by $\widetilde{Y}_t = Y_t - f(X_{t,1}, \ldots, X_{t,m})$ in (2.5). Therefore, they depend on the hyperparameters of the function f which are unknown in general. The use of (Y_t) does not cause additional problems since each maximum likelihood iteration is performed with a fixed $\widehat{\theta}_k$. Then, $Z_{1:T}|\{Y_t = y_t, t \in I_{av}\}, \widehat{\theta}_k$ has the same distribution as $Z_{1:T}|\{\widetilde{Y}_t = \widetilde{y}_t, t \in I_{av}\}, \widehat{\theta}_k$ (recall that the external regressors are known). Thus, if we need a sample from the former distribution, we can take a sample from the latter.

4.1 Approximation of the log-likelihood function based on the smoothing sample

As usual, the maximum likelihood strategy is to maximize the log-likelihood function $l(\theta)$ with respect to the unknown components of θ. Equivalently, we can maximize the difference $l(\theta) - l(\theta_k)$ for a given θ_k. The main idea of the following method is to approximate this difference using the particles $(z_{1:T}^{(j)})$ sampled according to the distribution of $Z_{1:T}|\{Y_t = y_t, t \in I_{av}\}, \theta_k$.

The first step is to rewrite the above difference using these smoothing particles. We find:

$$\begin{aligned}
&\delta l(\theta|\theta_k) := l(\theta) - l(\theta_k) \\
&= \log\left(\frac{p\left(\{y_t, t \in I_{av}\}|\theta\right)}{p\left(\{y_t, t \in I_{av}\}|\theta_k\right)}\right) \\
&= \log\left(\frac{\int \cdots \int p\left(\{y_t, t \in I_{av}\}, z_{1:T}|\theta\right) dz_1 \ldots dz_T}{p\left(\{y_t, t \in I_{av}\}|\theta_k\right)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \log\left(\int\cdots\int \frac{p\left(\{y_t, t\in I_{av}\}, z_{1:T}|\theta\right)}{p\left(\{y_t, t\in I_{av}\}, z_{1:T}|\theta_k\right)}\,\frac{p\left(\{y_t, t\in I_{av}\}, z_{1:T}|\theta_k\right)}{p\left(\{y_t, t\in I_{av}\}|\theta_k\right)} dz_1\ldots dz_T\right)\\
&= \log\left(\int\cdots\int \frac{p\left(\{y_t, t\in I_{av}\}, z_{1:T}|\theta\right)}{p\left(\{y_t, t\in I_{av}\}, z_{1:T}|\theta_k\right)} p\left(z_{1:T}|\{y_t, t\in I_{av}\}, \theta_k\right) dz_1\ldots dz_T\right)\\
&\approx \log\left(\frac{1}{N}\sum_{j=1}^{N}\frac{p\left(\{y_t, t\in I_{av}\}, z_{1:T}^{(j)}|\theta\right)}{p\left(\{y_t, t\in I_{av}\}, z_{1:T}^{(j)}|\theta_k\right)}\right)\\
&= \log\left(\frac{1}{N}\sum_{j=1}^{N}\frac{p\left(\{y_t, t\in I_{av}\}|z_{1:T}^{(j)}, \theta\right)}{p\left(\{y_t, t\in I_{av}\}|z_{1:T}^{(j)}, \theta_k\right)}\,\frac{p\left(z_{1:T}^{(j)}|\theta\right)}{p\left(z_{1:T}^{(j)}|\theta_k\right)}\right)\\
&= \log\left[\frac{1}{N}\sum_{j=1}^{N}\left(\prod_{t\in I_{av}}\frac{p\left(y_t|z_t^{(j)}, \theta\right)}{p\left(y_t|z_t^{(j)}, \theta_k\right)}\prod_{t=p+1}^{T}\frac{p\left(z_t^{(j)}|z_{(t-p):(t-1)}^{(j)}, \theta\right)}{p\left(z_t^{(j)}|z_{(t-p):(t-1)}^{(j)}, \theta_k\right)}\,\frac{p\left(z_{1:p}^{(j)}|\theta\right)}{p\left(z_{1:p}^{(j)}|\theta_k\right)}\right)\right]\\
&\approx \log\left[\frac{1}{N}\sum_{j=1}^{N}\left(\prod_{t\in I_{av}}\frac{p\left(y_t|z_t^{(j)}, \theta\right)}{p\left(y_t|z_t^{(j)}, \theta_k\right)}\prod_{t=p+1}^{T}\frac{p\left(z_t^{(j)}|z_{(t-p):(t-1)}^{(j)}, \theta\right)}{p\left(z_t^{(j)}|z_{(t-p):(t-1)}^{(j)}, \theta_k\right)}\right)\right].
\end{aligned}
\tag{4.1}
$$

Remark 4.1 *Some comments about the previous result.*
The sample $(z_{1:T}^{(j)})$ was generated using the smoothing algorithm described in Chapter 3. In fact, $(z_{1:T}^{(j)})$ is a sample of $Z_{1:T}|\{\widetilde{Y}_t = \widetilde{y}_t, t\in I_{av}\}, \theta_k$ and the latter random variable has the same distribution as $Z_{1:T}|\{Y_t = y_t, t\in I_{av}\}, \theta_k$.
Moreover, the densities $p\left(z_{1:p}^{(j)}|\theta\right)$ and $p\left(z_{1:p}^{(j)}, \theta_k\right)$ are multivariate normal densities. The expected values are given by the p-dimensional vector $(0,\ldots,0)$ and the covariance matrices can be computed using the Yule-Walker equations given θ and θ_k, respectively. Of course, it is possible to compute the ratio of these densities, but this evaluation requires some additional computations (solve the Yule-Walker equation systems to find the covariance matrices and compute the multivariate normal densities). Since for $T \gg p$ this ratio does not have a relevant effect on the result, we avoid these computations and we make an additional approximation to find (4.1).

Now, we examine closer the approximation (4.1) and we protect it against numeric overflow. To this end,

Definition 4.2 *Let*

$$s_j = \sum_{t \in I_{av}} \log \left(\frac{p\left(y_t | z_t^{(j)}, \theta\right)}{p\left(y_t | z_t^{(j)}, \theta_k\right)} \right) + \sum_{t=p+1}^{T} \log \left(\frac{p\left(z_t^{(j)} | z_{(t-p):(t-1)}^{(j)}, \theta\right)}{p\left(z_t^{(j)} | z_{(t-p):(t-1)}^{(j)}, \theta_k\right)} \right)$$

and $\quad s_{max} = \max_j s_j$.

We note that the variables s_j are sums of logarithmic or quadratic terms for the considered model. Thus, their computation is reliable and it has a complexity of order T.

We can estimate $\delta l(\theta|\theta_k)$ using (4.1) and the previous definition as follows:

$$\begin{aligned} \widehat{\delta l}(\theta|\theta_k) &= \log \left[\frac{1}{N} \sum_{j=1}^{N} \exp\left(s_j\right) \right] \\ &= \log \left[\exp\left(s_{max}\right) \cdot \sum_{j=1}^{N} \exp\left(s_j - s_{max}\right) \right] - \log(N) \\ &= s_{max} + \log \left[\sum_{j=1}^{N} \exp\left(s_j - s_{max}\right) \right] - \log(N). \qquad (4.2) \end{aligned}$$

The logarithm term in the previous expression does not cause numerical problems since its argument is between 1 and N. It is also possible to compute it accurately. Therefore, the leading term in the expression (4.2) is s_{max}. Since the computation of the s_j's is not problematic as noted before, $\widehat{\delta l}(\theta|\theta_k)$ is protected against numeric overflow. Moreover, (4.2) has a complexity of order TN.

Finally, the iterative estimation method based on the smoothing sample can be summarized as follows. It is initialized with a reliable estimate $\widehat{\theta}_0$ and it is carried on by

$$\widehat{\theta}_{k+1} = \arg\max_{\theta} \widehat{\delta l}(\theta|\widehat{\theta}_k)$$

until the differences $\widehat{\theta}_{k+1} - \widehat{\theta}_k$ fluctuate around zero. Note that the maximum is computed with respect to the unknown components of θ, and then $\widehat{\theta}_{k+1}$ is found by putting the new estimates and the fixed parameters together. In addition, enlarging the sample size N during the iterations reduces the Monte Carlo error in (4.2). This leads to a more precise estimate of θ.

4.2 MCEM method to approximate the log-likelihood function

The second method proposed to perform the parameter estimation is derived from the Expectation-Maximization (EM) algorithm.
The EM procedure is a quite general optimization method involving unobserved data, see for example Dempster et al. (1977) and McLachlan and Krishnan (1997). The key idea is to maximize the expected value of the log-likelihood of both observed and unobserved data. Since the expected value also depends on the parameters, the procedure is iterative. In addition, the expectation can be approximated using the Monte Carlo method which uses samples generated for a given set of parameters. This gives rise to the Monte Carlo EM (MCEM) algorithm, see for example Wei and Tanner (1990) and Chan and Ledolter (1995).

Explicitly, the first step is to compute the full log-likelihood function of the data $Z_{1:T}, \{Y_t, t \in I_{av}\}$. We find using the dependence structures of the model (2.1) and (2.2):

$$\begin{aligned}
&l(\theta|\{y_t, t \in I_{av}\}, z_{1:T}) = \log\left(p\left(\{y_t, t \in I_{av}\}, z_{1:T}|\theta\right)\right) \\
&= \log\left(p\left(\{y_t, t \in I_{av}\}|z_{1:T}, \theta\right)\ p\left(z_{1:T}|\theta\right)\right) \\
&= \sum_{t \in I_{av}} \log\left(p\left(y_t|z_t, \theta\right)\right) + \sum_{t=p+1}^{T} \log\left(p\left(z_t|z_{(t-p):(t-1)}, \theta\right)\right) + \log\left(p\left(z_{1:p}|\theta\right)\right).
\end{aligned}$$

The new objective function is defined by taking the expectation with respect to $Z_{1:T}|\{Y_t = y_t, t \in I_{av}\}, \theta_k$. Then, the expectation is approximated using the Monte Carlo method. We have,

$$\begin{aligned}
&Q(\theta|\theta_k) := \mathbf{E}_{\theta_k}\left[l(\theta|\{y_t, t \in I_{av}\}, z_{1:T}) - \log\left(p\left(z_{1:p}|\theta\right)\right)\right] \\
&= \mathbf{E}_{\theta_k}\left[\sum_{t \in I_{av}} \log\left(p\left(y_t|z_t, \theta\right)\right)\right] + \mathbf{E}_{\theta_k}\left[\sum_{t=p+1}^{T} \log\left(p\left(z_t|z_{(t-p):(t-1)}, \theta\right)\right)\right] \\
&\approx \frac{1}{N}\sum_{j=1}^{N}\sum_{t \in I_{av}}\log\left(p\left(y_t|z_t^{(j)}, \theta\right)\right) + \frac{1}{N}\sum_{j=1}^{N}\sum_{t=p+1}^{T}\log\left(p\left(z_t^{(j)}|z_{(t-p):(t-1)}^{(j)}, \theta\right)\right) \\
&=: \widehat{Q}(\theta|\theta_k).
\end{aligned}$$

$(z_{1:T}^{(j)})$ is a sample of $Z_{1:T}|\{Y_t = y_t, t \in I_{av}\}, \theta_k$. As noted in Section 4.1, these particles are generated using the smoothing algorithm of Chapter

3. Moreover, note that the computation of $\widehat{Q}(\theta|\theta_k)$ has a complexity of order TN.

Finally, the MCEM method can be summarized as follows. It is initialized with a reliable estimate $\widehat{\theta}_0$ and it is carried on by

$$\widehat{\theta}_{k+1} = \arg\max_{\theta} \widehat{Q}(\theta|\widehat{\theta}_k)$$

until the differences $\widehat{\theta}_{k+1} - \widehat{\theta}_k$ fluctuate around zero. As before, note that the maximum is computed with respect to the unknown components of θ, and then $\widehat{\theta}_{k+1}$ is found by putting the new estimates and the fixed parameters together.

The replacement of the expectation with the Monte Carlo expectation has the consequence that the typical monotonicity $l(\widehat{\theta}_{k+1}) \geq l(\widehat{\theta}_k)$ of the EM-sequence $(\widehat{\theta}_k)$ is lost. But, as already noted, the estimates $\widehat{\theta}_k$ fluctuate randomly around the true maximum likelihood estimate $\widehat{\theta}_{ML}$ after some iterations. The amplitude of the fluctuations is small already with moderate values of N. But the convergence may be slow, and thus it is important to begin the iterations with a good $\widehat{\theta}_0$.

4.3 Algorithm to compute starting estimates

We saw in the previous sections that it is important to begin the maximum likelihood recursion with a good estimate $\widehat{\theta}_0$. In fact, a good choice reduces the number of iterations needed and the estimation procedure is faster. Thus, the aim of this section is to develop an algorithm to compute reliable starting parameters for the model (2.1) and (2.2). It is convenient to separate the estimation of the hyperparameters of the function f from the estimation of the nuisance parameters. We will estimate first the hyperparameters and then the nuisance parameters. In both cases, the estimation procedures should be robust and also work in the presence of missing values in the observation series or in the external regressors.

If the model has external regressors, we first estimate the hyperparameters. Otherwise, we can skip this step and go directly to Subsection 4.3.2.

4.3.1 Estimation of hyperparameters

We have already remarked that the function f in (2.2) may be nonlinear. Of course, the situation is simpler when f is linear. In such a case, the hyperparameters can be estimated by applying the function *rlm* in the statistical software *R*, see R Development Core Team (2005). This function fits a robust linear model by iterated re-weighted least squares. It supplies different ψ functions to perform the robust fit: the Huber, Tukey bisquare and Hampel proposals. Moreover, it has methods to supply the estimates needed to start the re-weighted least squares iterations and it can work in the presence of missing values in (Y_t) or in the external regressors.

When f is nonlinear, we can generalize this technique by substituting the iterated re-weighted least squares method by a iterated re-weighted nonlinear least squares one. The nonlinear least squares fit can be achieved with the function *nls* in *R*. The robustness can be obtained using different ψ functions to compute the weights. The supplied ψ proposals are the Huber, Tukey bisquare and Hampel as in the linear case.

The function *nls* can work in the presence of missing values in (Y_t) or in the external regressors. But, it needs starting estimates. We can find them by attempts or from previous studies. In fact, no automatic procedure is known to compute starting values in the general nonlinear regression problem.

In most cases, we use the Tukey bisquare proposal to find the robust hyperparameter estimates for both a linear or nonlinear function f. Once the estimates are found, we can compute the model residuals which are the input series to estimate the nuisance parameters.

4.3.2 Estimation of nuisance parameters

If the hyperparameters have already been estimated or if there are no external regressors, the considered model can be simplified to (2.6) and (2.7).

We assume for the moment that all nuisance parameters are unknown. The key ideas to estimate them are based on both the Yule-Walker equations for the AR process (Z_t) and the relationship between the autocovariance function of (Z_t) and the corresponding one of (Y_t).

Definition 4.3 *For any integer h define*

$$\begin{aligned}
C(h) &= Cov\left(\widetilde{Y}_t, \widetilde{Y}_{t-h}\right),\\
\gamma(h) &= Cov\,(Z_t, Z_{t-h}),\\
\rho(h) &= Cor\,(Z_t, Z_{t-h}),\\
\rho_{part}(h) &= Cor\,\left(Z_t, Z_{t-h}|Z_{(t-h+1):(t-1)} = z_{(t-h+1):(t-1)}\right),\\
\sigma_\varepsilon^2 &= Var\,(\varepsilon_t).
\end{aligned}$$

Then, it follows directly from the i.i.d. assumption on the observation errors (ε_t) that

$$\begin{aligned}
C(0) &= Cov\,(Z_t + \varepsilon_t, Z_t + \varepsilon_t) = \gamma(0) + \sigma_\varepsilon^2,\\
C(h) &= Cov\,(Z_t + \varepsilon_t, Z_{t-h} + \varepsilon_{t-h}) = \gamma(h), \qquad h = 1, \dots, p+1.
\end{aligned}$$

In addition, it follows from the Yule-Walker equations for (Z_t) that

$$\begin{aligned}
\gamma(h) &= \sum_{l=1}^{p} \varphi_l \gamma(h-l), \qquad h = 1, \dots, p+1,\\
\gamma(0) &= \sum_{l=1}^{p} \varphi_l \gamma(l) + \sigma_V^2.
\end{aligned}$$

If we combine the last four (systems of) equalities and we assume that estimates of $C(0)$,..., $C(p+1)$ have been computed, then we have a system of $p+2$ equations for the $p+2$ unknown parameters φ_1, ..., φ_p, σ_V and σ_ε (recall that $\gamma(-h) = \gamma(h)$). Thus, in principle, we could estimate these parameters by solving the system. But it is difficult to solve it in closed form already for an AR(2) state process, and existence and uniqueness of solutions in the stationarity region are not clear. The equation system needs to be solved numerically. We note that

$$\gamma(0) = C(0) - \sigma_\varepsilon^2, \tag{4.3}$$

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{C(h)}{C(0) - \sigma_\varepsilon^2} = \frac{C(h)}{C(0)}\,(1+\eta), \quad h = 1, \dots, p+1, \tag{4.4}$$

$$\sigma_V^2 = \gamma(0) - \sum_{l=1}^{p} \varphi_l \gamma(l) \tag{4.5}$$

with

$$\eta = \frac{\sigma_\varepsilon^2}{\gamma(0)} > 0. \tag{4.6}$$

Thus, if $C(0), \ldots, C(p+1)$ have already been estimated, we can solve the equation system (4.4) numerically with respect to the unknown parameters $\varphi_1, \ldots, \varphi_p$ and η (recall that the autocorrelations $\rho(h)$ depend on $\varphi_1, \ldots, \varphi_p$). Then, the estimates of σ_V and σ_ε can be found from (4.3), (4.5) and (4.6).

How do we succeed in solving (4.4) numerically? We use the following well-known property of the partial autocorrelations of a stationary AR(p) process (Z_t):

$$\begin{aligned} |\rho_{part}(h)| &< 1 \qquad \forall\, h = 1, \ldots, p, \\ \rho_{part}(p+1) &= 0. \end{aligned}$$

Thus, we set

Definition 4.4 *Define*

$\eta_{max} = \sup\{\eta \geq 0 \,|\,$ *the vector* $(\rho_\eta(h))$, $h = 1, \ldots, p$, *is in the stationarity region of the AR(p) process*$\}$
with $\rho_\eta(h) = \frac{C(h)}{C(0)}\,(1+\eta)$ *as in (4.4).*
In addition, define an equally spaced grid $\eta_k = (k-1)\Delta\eta$, $k = 1, \ldots, N(\eta_{max})$, $\Delta\eta$ *given.* $N(\eta_{max})$ *is the smallest integer not less than* $\eta_{max}/\Delta\eta$.

Now, the strategy is to compute for each grid point η_l the corresponding autocorrelations $\rho_{\eta_l}(h)$, $h = 1, \ldots, p+1$, and find the AR($p+1$) process which has exactly these autocorrelations. This can be achieved easily by applying the Durbin-Levinson algorithm, a recursive procedure for calculating the AR parameters $\varphi(\eta_l)_j$ from the autocorrelations $\rho_{\eta_l}(h)$ $(j, h = 1, \ldots, p+1)$. It has the nice feature that the partial autocorrelations are also found during the recursion since $\rho_{part}(m) = \varphi_m$ in an AR(m) model. The recursion and more details can be found for example in Box et al. (1994), Appendix A 3.2.
Once the AR($p+1$) models have been found for all grid points, we select the AR($p+1$) model which is nearest to the AR(p) we are looking for. To this end, we introduce

Definition 4.5 *Let*

$$\widehat{\eta} = \arg\min_{\eta_l} |\rho_{part,\eta_l}(p+1)| = \arg\min_{\eta_l} \left|\varphi(\eta_l)_{p+1}\right|.$$

Then, we consider the AR(p) model with coefficients $\widehat{\varphi}_j = \varphi(\widehat{\eta})_j$, $j = 1, \ldots, p$, as approximation for the AR(p) process which solves the

equation system (4.4).
Sometimes it happens that the selected $\widehat{\varphi}_1, \ldots, \widehat{\varphi}_p$ give rise to a non-stationary AR(p) process. For example, this may happen when the true state process is near the stationarity border, since the Durbin-Levinson algorithm is very sensitive to rounding errors in this situation. What we do in this circumstance is to enforce stationarity by moving the roots of the characteristic polynomial which are inside or on the boundary of the unit circle just outside it. The translation is carried out along the radius of the unit circle. Then, the new stationary AR(p) process with these roots is computed.

Now, it is possible to estimate σ_V and σ_ε. From (4.3) and (4.6) we get

$$\widehat{\sigma}_\varepsilon = \sqrt{\frac{\widehat{\eta}}{1+\widehat{\eta}} C(0)} \tag{4.7}$$

and the auxiliary result

$$\widehat{\gamma(0)} = \frac{C(0)}{1+\widehat{\eta}}.$$

Therefore, it follows from the auxiliary result and (4.5)

$$\widehat{\sigma}_V = \sqrt{\frac{C(0)}{1+\widehat{\eta}} - \sum_{l=1}^{p} \widehat{\varphi}_l \; C(l)}. \tag{4.8}$$

Sometimes it occurs that the argument of the square roots in (4.7) or (4.8) is non-positive. For example, this may happen when the true unknown value of σ_V or σ_ε is near zero. Or $\widehat{\eta}$ can be zero. This can occur when the time series is too short. In these situations, we set the corresponding estimate equal to a small positive value.
In addition, it follows from Definition (4.3) and the result (2.4) that

$$\sigma_\varepsilon^2 = Var\left(\varepsilon_t\right) = \frac{c^2}{2m-3}.$$

In principle, we would need an additional equation to estimate c and m. But the estimation of the degrees of freedom m is known to be complex. Therefore, we adopt a pragmatic approach: m is not estimated directly but several discrete values for $\widehat{m}$ are chosen, typically $\widehat{m} \in \{1,2,3,4,5\}$. Consequently, we have several starting values for $\widehat{c}$:

$$\widehat{c} = \sqrt{\max(1, 2\widehat{m} - 3)}\ \ \widehat{\sigma}_\varepsilon .$$

Thus, we will use all these different starting values to compute the estimates using the methods described in Sections 4.1 or 4.2. Between these estimates, the estimate $\widehat{\theta}$ which maximizes the log-likelihood over the different starting values of $\widehat{m}$ is chosen as final estimate of the unknown coefficients.
Note that small values of $\widehat{m}$ (1 or 2) lead to very robust filtering and smoothing recursions.

Four questions are still open: how we can find η_{max} and the estimates of $C(0), \ldots, C(p+1)$ and what we do if some nuisance parameters are known or there are missing values in the series $(\widetilde{Y}_t)$.
The value of η_{max} can be found easily if the state equation is an AR(1) or an AR(2) process. In fact, the stationarity region of an AR(1) process is given by

$$1 \ > \ |\rho_{part,\eta}(1)| = |\rho_\eta(1)| = \left|\frac{C(1)}{C(0)}\right| \ (1+\eta)$$

which results in the condition

$$\eta_{max} = \left|\frac{C(0)}{C(1)}\right| - 1. \tag{4.9}$$

The stationarity region of an AR(2) process can be derived as in Box et al. (1994), Subsection 3.2.4. It is given by

$$\rho_\eta(2) < 1 \quad \text{ and } \quad \rho_\eta(2) > 2\rho_\eta(1)^2 - 1.$$

These conditions correspond to the area in the $(\rho(1), \rho(2))$-plane bounded below by the parabola $\rho(2) = 2\rho(1)^2 - 1$ and above by the horizontal line $\rho(2) = 1$. Since $\rho_\eta(2)/\rho_\eta(1) = C(2)/C(1)$, we deduce that η_{max} is attained on the upper border $\rho(2) = 1$ for $C(2) \geq |C(1)|$. Otherwise, it lies on the parabola $\rho(2) = 2\rho(1)^2 - 1$. Explicitly,

- If $C(2) \geq |C(1)|$: from

$$1 = \rho_\eta(2) = \frac{C(2)}{C(0)}\ (1+\eta),$$

it follows that

$$\eta_{max} = \frac{C(0)}{C(2)} - 1. \tag{4.10}$$

- If $C(2) \leq |C(1)|$: from

$$\rho_\eta(2) = 2\rho_\eta(1)^2 - 1,$$

it follows that

$$\begin{aligned}\frac{C(2)}{C(0)}\ (1+\eta) &= 2\ \frac{C(1)^2}{C(0)^2}\ (1+\eta)^2 - 1,\\ 0 &= 2\ C(1)^2\ (1+\eta)^2 - C(0)\ C(2)\ (1+\eta) - C(0)^2,\\ (1+\eta)_\pm &= \frac{C(0)\ C(2) \pm \sqrt{C(0)^2\ C(2)^2 + 8\ C(1)^2\ C(0)^2}}{4\ C(1)^2}.\end{aligned}$$

The solution with the minus sign implies that $1+\eta \leq 0$ which is not acceptable (η cannot be negative). Therefore,

$$\eta_{max} = \frac{C(0)\ C(2) + \sqrt{C(0)^2\ C(2)^2 + 8\ C(1)^2\ C(0)^2}}{4\ C(1)^2} - 1. \tag{4.11}$$

The calculation of the stationarity region of an AR(p) process, $p \geq 3$, is no more straightforward. But, since the stationarity regions are nested, we propose to choose η_{max} as in (4.10) or (4.11) also for these processes. As a consequence of this nearly optimal choice, we should expect that all grid points are outside the stationarity region from one $\eta_{l'}$. This will be reflected in the Durbin-Levinson algorithm by some terms $\left|\rho_{part,\eta_{l'}}(h)\right|$, $h = 1, \ldots, p$, which exceed 1 from an index h'. In these cases, we can stop the Durbin-Levinson procedure for $\eta_{l'}$ and also for the following grid points.

The second open question is the estimation of the autocovariances $C(0)$,..., $C(p+1)$. They have to be estimated carefully, since they play a crucial role in the estimation of the nuisance parameters as we saw above. Especially, the estimation procedure should be robust.

The main idea is to approximate the observation series $(\widetilde{Y}_t)$ by a Gaussian AR($p+1$) process:

$$\widetilde{Y}_t = \phi_1 \widetilde{Y}_{t-1} + \cdots + \phi_{p+1}\widetilde{Y}_{t-p-1} + \widetilde{V}_t, \qquad \widetilde{V}_t \sim \mathcal{N}(0, \sigma^2_{\widetilde{Y}}). \tag{4.12}$$

Now, the AR($p+1$) process can be interpreted as a regression of $\widetilde{Y}_t$ on $\widetilde{Y}_{t-1}$, ..., $\widetilde{Y}_{t-p-1}$. Therefore, robust estimates $\widehat{\phi}_1$, ..., $\widehat{\phi}_{p+1}$ and $\widehat{\sigma}_{\widetilde{Y}}$

can be found by fitting a robust regression. To this end, we use again the function *rlm* with the Tukey bisquare proposal. Recall that the function can work in the presence of missing values in $(\widetilde{Y}_t)$. Since the AR$(p+1)$ process is only an approximation of the series $(\widetilde{Y}_t)$, it could happen that the estimates $\widehat{\phi}_1$, ..., $\widehat{\phi}_{p+1}$ give rise to a non-stationary AR$(p+1)$ process. In these cases, we enforce stationarity as explained before.
Now, the Yule-Walker equations for the AR$(p+1)$ approximation of $(\widetilde{Y}_t)$ yield

$$C(h) = \sum_{l=1}^{p+1} \phi_l C(h-l), \qquad h = 1, \ldots, p+1 \tag{4.13}$$

and

$$C(0) = \sum_{l=1}^{p+1} \phi_l C(l) + \sigma^2_{\widetilde{Y}}. \tag{4.14}$$

If we replace the true parameters ϕ_1, ..., ϕ_{p+1} and $\sigma_{\widetilde{Y}}$ with the estimated ones, we get a system of $p+2$ linear equations for the unknown autocovariances $C(0)$,..., $C(p+1)$. It is easy to solve this system numerically.

Finally, if some nuisance parameters are known, we replace the estimated values with these given values, paying attention to preserve the stationarity. In addition, we note that the presence of missing values in $(\widetilde{Y}_t)$ does not cause problems. In fact, the series $(\widetilde{Y}_t)$ is only used to compute the robust linear regression with the function *rlm*, and this function can work in the presence of missing values.

It is useful to recapitulate the whole procedure to estimate the starting nuisance parameters.

Algorithm 4.1 *Starting estimates of the nuisance parameters.*
Assumption:
The hyperparameters have already been estimated or there are no external regressors. Therefore, the model is given by (2.6) and (2.7).

Then:

1. *Compute the robust regression of $\widetilde{Y}_t$ on $\widetilde{Y}_{t-1}$, ..., $\widetilde{Y}_{t-p-1}$ using the function* rlm *with the Tukey bisquare proposal. This function*

can work in the presence of missing values in $(\widetilde{Y}_t)$.
In this way, robust estimates $\widehat{\phi}_1, \ldots, \widehat{\phi}_{p+1}$ *and* $\widehat{\sigma}_{\widetilde{Y}}$ *are found, see (4.12). If they give rise to a non-stationary AR*($p+1$) *process, transform the coefficients* $\widehat{\phi}_1, \ldots, \widehat{\phi}_{p+1}$*to enforce stationarity.*

2. *Substitute the previous estimates in the system of* $p+2$ *linear equations (4.13) and (4.14). Solve this system numerically with respect to* $C(0), \ldots, C(p+1)$.
3. *Substitute the robust estimates* $\widehat{C(0)}, \ldots, \widehat{C(p+1)}$ *in (4.3), (4.4), (4.5) and (4.6) and estimate the nuisance parameters.*

   To this end:

   - *Find* η_{max}.
     For an AR(1) *process, it is given by (4.9). Otherwise by (4.10) or (4.11).*
   - *Define an equally spaced grid* $\eta_k = (k-1)\Delta\eta$, $k = 1, \ldots, N(\eta_{max})$, $\Delta\eta$ *given.* $N(\eta_{max})$ *is the smallest integer not less than* $\eta_{max}/\Delta\eta$.
   - *For each grid point* η_l, *estimate the AR coefficients of the AR*($p+1$) *model with autocorrelations given by* $\rho_{\eta_l}(h)$, $h = 1, \ldots, p+1$, *see (4.4). Use the Durbin-Levinson algorithm to estimate the coefficients. If* $\left|\rho_{part,\eta_{l'}}(h)\right|$ *exceeds* 1 *for a* $h = 1, \ldots, p$, *and a grid point* $\eta_{l'}$, *then stop the Durbin-Levinson procedure for this* $\eta_{l'}$ *and the successive ones since the autocorrelations will be outside the AR(p) stationarity region.*
   - *Select from the previous AR*($p+1$) *models the one which is nearest to an AR(p) process. I.e., let*

     $$\widehat{\eta} := \arg\min_{\eta_l} |\rho_{part,\eta_l}(p+1)| = \arg\min_{\eta_l} \left|\varphi(\eta_l)_{p+1}\right|$$

     and consider the AR(p) process with coefficients $\widehat{\varphi}_j = \varphi(\widehat{\eta})_j$, $j = 1, \ldots, p$.
     Check if this AR(p) process is stationary. If not, enforce stationarity.
   - *Find the estimates of the other nuisance parameters.*
     $\widehat{\sigma}_V$ *is given by (4.8). For* $\widehat{m}$, *choose several discrete values, typically* $\widehat{m} \in \{1,2,3,4,5\}$, *and compute the corresponding starting values for* $\widehat{c}$: $\widehat{c} = \sqrt{\max(1, 2\widehat{m}-3)}\,\widehat{\sigma}_\varepsilon$, *with* $\widehat{\sigma}_\varepsilon$ *as in (4.7).*

(Recall that we will use all these different starting values to compute the estimates using the methods described in Sections 4.1 or 4.2. Between these estimates, the estimate $\widehat{\theta}$ which maximizes the log-likelihood over the different starting values of $\widehat{m}$ is chosen as final estimate of the unknown coefficients.)

- *If some nuisance parameters are known, replace the estimated values with these given values. Pay attention to preserve the stationarity.*

Chapter 5

Examples

In this chapter, we illustrate the performances of the developed Monte Carlo algorithms with some examples. To this aim, the estimates obtained by the new algorithms are compared with the ones found using the Kalman recursions.
Before we begin, we recall that the Kalman filtering and smoothing recursions are defined as in Chapter 1. In addition, we compute the distribution of the stationary AR(p) process. This is a multivariate normal distribution and its covariance matrix is found using the Yule-Walker equations. This distribution is used as starting distribution for $Z_{(1-p):0}$ in the Kalman recursions and to generate the sample $(z^{(i)}_{(1-p):0})$ to begin the Monte Carlo algorithms.

The chapter is organized as follows. In the first two sections, we carry out simulation studies to compare the Kalman and the Monte Carlo algorithms. An example with real data is analysed in the third section.

5.1 State estimates given the parameters

In the simulation studies of this section, the parameters characterising the models are assumed to be known. The goal is to compare the estimates of the unknown state variables (Z_t) that are found with the developed Monte Carlo algorithms and the Kalman ones. The comparisons are done using a state space model with Gaussian error distributions where, first, we have no outliers, and then we consider an observation outlier at a fixed time point. For the sake of illustration, this state space model has no external regressors and the state equation is an AR(1) process. Thus, we write it using the same notation as in (2.6) and (2.7).

5.1.1 Gaussian error distributions without outliers

Two time series $\widetilde{Y}_{1:500}$ are simulated according to the model

$$Z_t = \varphi_1 Z_{t-1} + V_t, \qquad V_t \sim \mathcal{N}(0, \sigma_V^2), \tag{5.1}$$

$$\widetilde{Y}_t = Z_t + W_t, \qquad W_t \sim \mathcal{N}(0, \sigma_W^2) \tag{5.2}$$

with $\varphi_1 = 0.8$. In addition, we choose the values $\sigma_V = 1$, $\sigma_W = 1$ for the first time series and $\sigma_V = 4$, $\sigma_W = 1$ for the second one. The AR(1) state process is simulated with the *R* function *arima.sim* which uses a "burn-in" period at the beginning.

The next step consists of estimating the unknown state variables (Z_t) using the Kalman and the Monte Carlo algorithms, both using the filtering and the smoothing methods. The Kalman algorithms can be applied directly since the chosen error distributions are Gaussian, see (5.1) and (5.2). On the other hand, the new Monte Carlo algorithms require the choice of some additional parameters since the observation error distribution is supposed to be a Pearson type VII distribution, see the observation equation (2.7). Therefore, values for m and c are needed. For m, we choose two different values: $m = 1$ and $m = 3$. The aim is to compare two options which have different robust behaviours. This can be seen by noting that $m = 1$ and $m = 3$ will produce a t_1 and a t_5 densities, if $c = 1$ and $c = \sqrt{5}$, respectively (see Remark 2.1). In addition, c is chosen such that the Pearson observation error distribution assumed for the Monte Carlo algorithms and the $\mathcal{N}(0, \sigma_W^2)$ distribution used in the simulation have the same interquartile distance. This is done to make these distributions comparable. According to Remark 2.1, we set

$$c = \sqrt{2m-1}\, \sigma_W \, \frac{q_{\mathcal{N}(0,1)}(0.75)}{q_{t_{2m-1}}(0.75)}$$

where $q_{\mathcal{N}(0,1)}(0.75)$ is the 0.75 quantile of the $\mathcal{N}(0,1)$ distribution and $q_{t_{2m-1}}(0.75)$ is the 0.75 quantile of the t_{2m-1} one. Moreover, we analyse the influence of the sample size N on the results by taking two different values for it: $N = 200$ and $N = 1000$. Finally, the Monte Carlo smoothing algorithm is computed with both developed methods, see Sections 3.3 and 3.4. In both cases, the maximal allowed number of rejections before improving the distribution setup is 50.

Table 5.1 summarizes schematically all considered algorithms and options to estimate the unknown state variables (Z_t). It also contains the names given to these methods to permit easier references later. Furthermore, the table reports the CPU times measured on a dual-processor Intel Xeon 2.4 GHz (2000 MB RAM) to estimate each case once (this is the sum of the CPU times of the filtering and smoothing re-

Simulation		Estimation				
σ_V	σ_W	algorithm (smoothing method)	m	N	CPU time (in sec)	name
1	1	Kalman (-)	-	-	0.95	K a)
		Monte Carlo (1)	3	200	13.81	MC a)
		(1)		1000	207.44	MC b)
		(2)		200	5.19	MC c)
		(2)		1000	32.60	MC d)
		(1)	1	200	13.61	MC e)
		(1)		1000	208.50	MC f)
		(2)		200	5.28	MC g)
		(2)		1000	32.89	MC h)
4	1	Kalman (-)	-	-	1.16	K b)
		Monte Carlo (1)	3	200	12.96	MC i)
		(1)		1000	204.00	MC j)
		(2)		200	5.28	MC k)
		(2)		1000	33.83	MC l)
		(1)	1	200	14.65	MC m)
		(1)		1000	214.13	MC n)
		(2)		200	5.72	MC o)
		(2)		1000	37.14	MC p)

Table 5.1: *Simulation parameters and estimation methods in the exact Gaussian simulation study. Same parameters or algorithms are written only once to have a better readability. (1) or (2) denote the used Monte Carlo smoothing method for the estimation, see Sections 3.3 and 3.4, respectively.*

cursions). The Kalman algorithms are implemented with the statistical software *R*. The Monte Carlo algorithms are implemented principally in C, and *R* plays the role of the interface.
As expected, the Kalman recursions are faster than the Monte Carlo ones. There are noticeable gains in the CPU times used by the second Monte Carlo smoothing method with respect to the first one, although we have measured only one CPU time for each condition (recall that the filtering algorithm is the same in both cases). The CPU times of the second Monte Carlo smoothing algorithm are actually not bad considering also the length of the time series. The efforts to implement it are worthwhile.

As a second point, we examine the estimates of the unknown state variables (Z_t). We consider two quantities:

- $Z_t - med\left(Z_t|\widetilde{Y}_{1:t}\right)$ and $Z_t - med\left(Z_t|\widetilde{Y}_{1:T}\right)$:
  the bias between the true Z_t and the conditional median of Z_t computed using both the filtering and the smoothing recursions. For the Kalman algorithms, the medians are given by the conditional expected values. For the Monte Carlo algorithms, the medians are computed from the sampled particles.
- $\frac{\left|Z_t - med\left(Z_t|\widetilde{Y}_{1:t}\right)\right|}{sd\left(Z_t|\widetilde{Y}_{1:t}\right)}$ and $\frac{\left|Z_t - med\left(Z_t|\widetilde{Y}_{1:T}\right)\right|}{sd\left(Z_t|\widetilde{Y}_{1:T}\right)}$:
  the standardized absolute deviation between the true Z_t and the conditional median of Z_t computed using both the filtering and the smoothing recursions. For the Kalman algorithms, the denominators are given by the conditional standard deviations (square roots of the estimated variances). For the Monte Carlo algorithms, the denominators are computed using the median absolute deviations (MAD) of the sampled particles adjusted by a factor to have consistency at the normal distribution.

The results are shown in Figures 5.1, 5.2, 5.3 and 5.4 for the two considered time series. The names are as in Table 5.1. The box plots computed with the first Monte Carlo smoothing method are omitted for the clarity. In fact, they present the same characteristics as the ones computed with the second method since the smoothing target density is the same and only the sampling technique is different. The filtering recursion is the same in both cases and therefore the filtering samples are identical.
The Monte Carlo box plots with $m = 3$ have the same shape as the

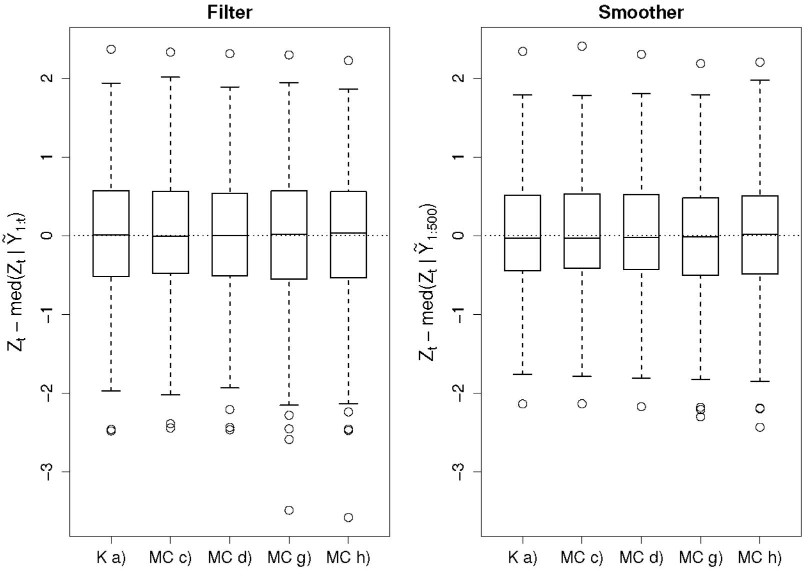

Figure 5.1: *Bias for different estimation methods in the exact Gaussian simulation study. The time series is simulated with $\varphi_1 = 0.8$, $\sigma_V = 1$ and $\sigma_W = 1$. The methods' names are explained in Table 5.1.*

Kalman ones. The Monte Carlo box plots with $m = 1$ have a little bit wider interquartile distance and they present more outliers. This behaviour is not surprising since the Monte Carlo filtering and smoothing algorithms with $m = 1$ are more robust than the ones with $m = 3$. Thus, extreme observations caused by extreme (unknown) states Z_t may be interpreted wrongly as outliers (see the circles in the box plots).

In all box plots, it is very common to find that the first observations are treated as outliers. This is to be expected. The Kalman algorithms are started with given values for the expectation and the variance. Similarly, the Monte Carlo algorithms are initialized with a sample $(z_0^{(i)})$ from the true distribution of Z_0. It takes some time steps until these initial choices lose their effect. But not too long, as we can see.

In addition, we note that in some cases the box plots computed with the smoothing algorithms are a little bit more concentrated than the ones computed with the filtering algorithms. This is to be expected since the smoothing reconstructions profit from all observations in contrast to the filtering ones.

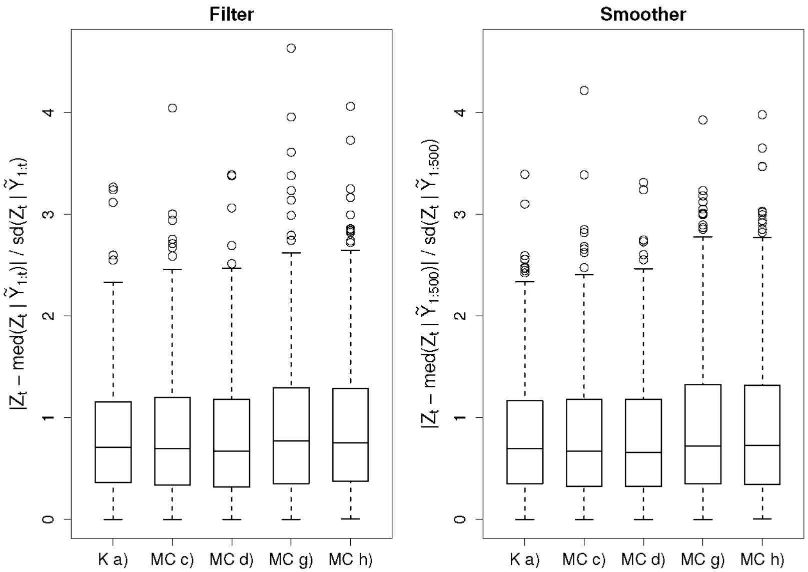

Figure 5.2: *Standardized absolute deviation for different estimation methods in the exact Gaussian simulation study. The time series is simulated with $\varphi_1 = 0.8$, $\sigma_V = 1$ and $\sigma_W = 1$. The methods' names are explained in Table 5.1.*

The Monte Carlo sample size N does not seem to affect the box plots. The final results are already good with the quite small size $N = 200$.

The next point we discuss using this simulation study is the sampling efficiency of the computed Monte Carlo methods. An intuitive quantity to measure it is given by the ratio of the number of rejections to produce the samples, divided by the sample size. The results are shown in Figures 5.5 and 5.6.

A preliminary remark before we discuss the results. The smoothing particles for $T = 500$ are given by the filtering particles (recall that 500 is the length of the simulated time series). This permits to initialize the smoothing recursions. But, as a consequence, there are no rejections at $T = 500$, and we omit this time point from the box plots.

As we can see, some characteristics are the same for the two considered time series. The ratios are very small in the filtering algorithm for both $m = 1$ and $m = 3$. Thus, the sampling efficiencies are very high. The cases with $m = 1$ present more outliers. Again, this feature can be

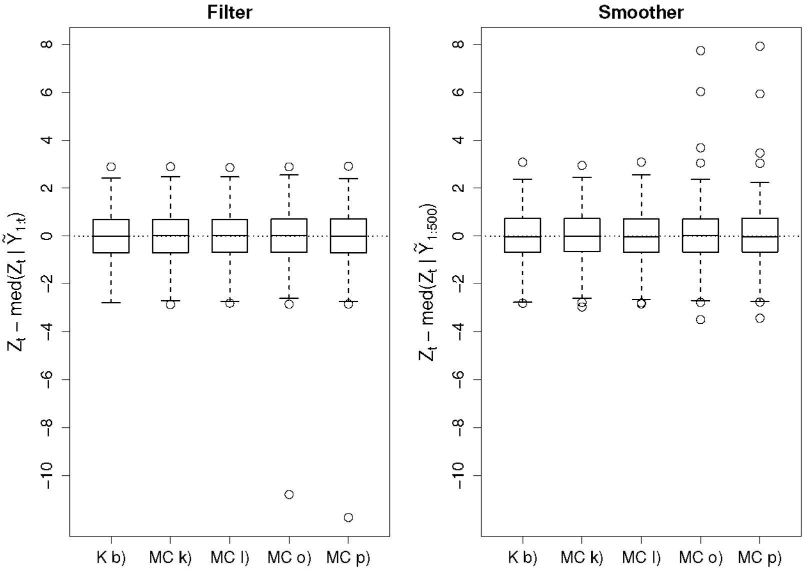

Figure 5.3: *Same as Figure 5.1, with* $\sigma_V = 4$.

explained by the higher robustness of this choice. In addition, a larger sample size should reduce the standard deviation of the ratios without changing the expected value. This is also noticeable, although not really well for $m = 1$ in Figure 5.6.

The rejection ratio is very stable with respect to the different values of m and N in the second Monte Carlo smoothing method. This is a consequence of the used procedure to implement it. In fact, each particle is sampled separately from the others, and therefore the proposal setup can be adapted to each particle. Unfortunately, the ratio is greater than in the filtering algorithm. This is not a surprise since the smoothing recursion is more complex than the filtering one as we discussed in Chapter 3. But the ratio is still acceptable and definitely better than with the first smoothing method. In fact, in the latter method, the same proposal setup is used to sample more particles and it is improved only when it becomes too bad. This happens when the number of rejections exceeds a chosen bound.

It is interesting to note that with the first smoothing algorithm, the ratio exhibits a similar behaviour for the two different degrees of freedom ($m = 1$ and $m = 3$) but not for the two sample sizes ($N = 200$ and

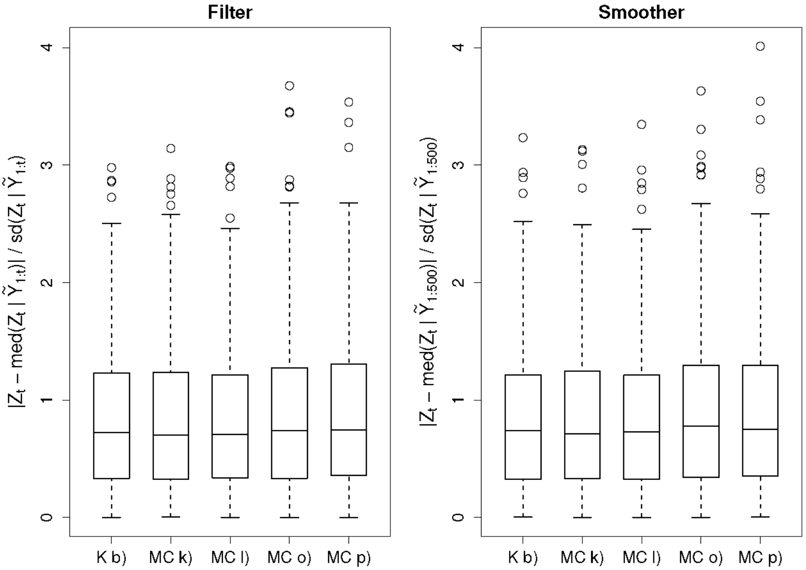

Figure 5.4: *Same as Figure 5.2, with $\sigma_V = 4$.*

$N = 1000$). This can be explained looking at the number of computed proposal setups at each time t, $t = 1, \ldots, 500$. Its summary is reported in Table 5.2 for the different methods. On the whole, the number of computed setups is stable for different values of m and N. But, if the sample size N is 1000, the maximal allowed number of rejections may

	MC a)	MC b)	MC e)	MC f)	MC i)	MC j)	MC m)	MC n)
Min.	2	3	2	3	2	2	2	2
1st Qu.	3	3	3	4	2	3	3	4
Median	3	4	3	4	2	3	3	5
Mean	2.99	3.61	3.15	3.92	2.33	3.12	3.47	4.60
3rd Qu.	3	4	3	4	3	3	4	5
Max.	4	5	5	6	4	5	5	7

Table 5.2: *Summary of the number of computed setups for different Monte Carlo smoothing methods in the exact Gaussian simulation study. The smoothing recursions are performed with the first idea. The methods' names are explained in Table 5.1*

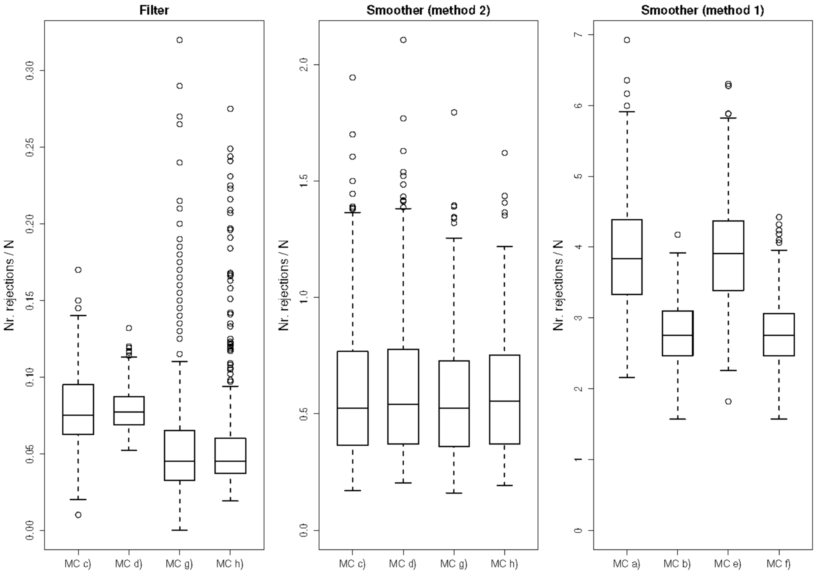

Figure 5.5: *Rejection ratios for different Monte Carlo estimation methods in the exact Gaussian simulation study. The time series is simulated with $\varphi_1 = 0.8$, $\sigma_V = 1$ and $\sigma_W = 1$. The methods' names are explained in Table 5.1*

be reached more often than with $N = 200$. This results in a slightly greater number of computed setups for $N = 1000$ than for $N = 200$. Consequently, the used setups for $N = 1000$ are somewhat more accurate and, combined with the fact that the particles are more dense for $N = 1000$, it leads to a lower rejection ratio for this case.

Finally, Table 5.3 reports the total number of setup improvements in the Monte Carlo smoothing recursions computed with the second idea. All times t, $t = 1, \ldots, 500$, are taken together since the number of setup improvements is never greater than 1 for each single time t. We can see that the value of m influences the total number of computed setups. As before, the reason is the higher robustness of the algorithm with $m = 1$. In addition, the value of N combined with $m = 1$ also influences this number. The explanation can be as discussed above for the influence of N on the rejection ratio in the Monte Carlo smoothing algorithm computed with the first method. But, on the whole, the total number of computed setups is small. This feature also contributes to make the

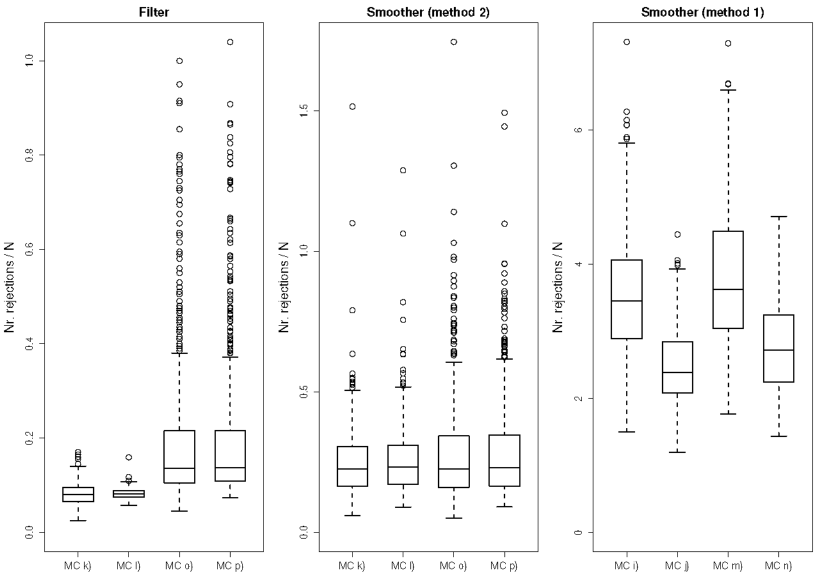

Figure 5.6: *Same as Figure 5.5, with $\sigma_V = 4$.*

MC c)	MC d)	MC g)	MC h)	MC k)	MC l)	MC o)	MC p)
3	2	103	139	5	7	64	121

Table 5.3: *Total number of computed setups for different Monte Carlo smoothing methods in the exact Gaussian simulation study. The smoothing recursions are performed with the second idea. The methods' names are explained in Table 5.1*

second smoothing method faster than the first one.

5.1.2 Gaussian error distributions with one outlier

In this subsection, the influence of a isolated outlier at a fixed time point is analyzed using a Gaussian state space model with known parameters. For the purpose of illustration, the simple state space model considered in Subsection 5.1.1 is also used here. The analysis is repeated for two simulated time series with parameters given by $\varphi_1 = 0.2, \sigma_V = 4, \sigma_W = 1$ and $\varphi_1 = 0.8, \sigma_V = 1, \sigma_W = 1,$ respectively. The time series have length 100. In addition, the outlier is at the fixed time $t = 60$ for both series. Thus, we vary the value of $\widetilde{y}_{60}$ and we estimate the state Z_{60} using different methods. In this way, we can examine the behaviour of the estimated state Z_{60} as function of $\widetilde{y}_{60}$ under different methods. These methods are the Kalman algorithms and the Monte Carlo ones with different options (in all cases the estimates are computed using both the filtering and smoothing recursions).

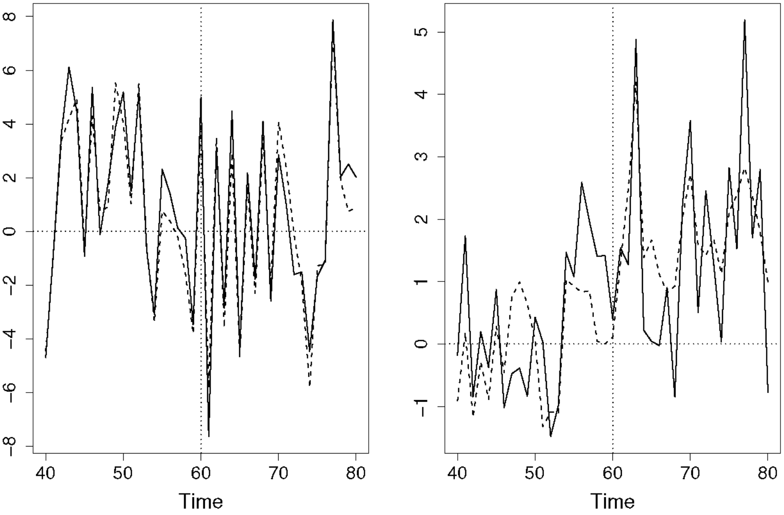

Figure 5.7: *Section of the two considered time series in the simulation study with one outlier. On the left, the time series with $\varphi_1 = 0.2$, $\sigma_V = 4$, $\sigma_W = 1$; on the right, the series with $\varphi_1 = 0.8$, $\sigma_V = 1$, $\sigma_W = 1$. The solid line gives $(\widetilde{Y}_t)$, the dashed line (Z_t), and the vertical dotted line the outlier location.*

We start with the first time series. A section of it around $t = 60$ is shown in Figure 5.7. The state estimates with the Kalman recursions can be computed directly. For the Monte Carlo recursions, we choose $m = 3$, $N = 2000$ and we use the second smoothing method.

Figure 5.8 shows the median, the 10% and 90% quantiles of the state estimates of Z_{60} as function of $\widetilde{y}_{60}$ for both the Kalman and the Monte Carlo recursions. The filtering results are on the left, the smoothing ones on the right. In addition, the vertical dashed line gives the location of the original simulated $\widetilde{y}_{60}$ and the horizontal dashed line corresponds

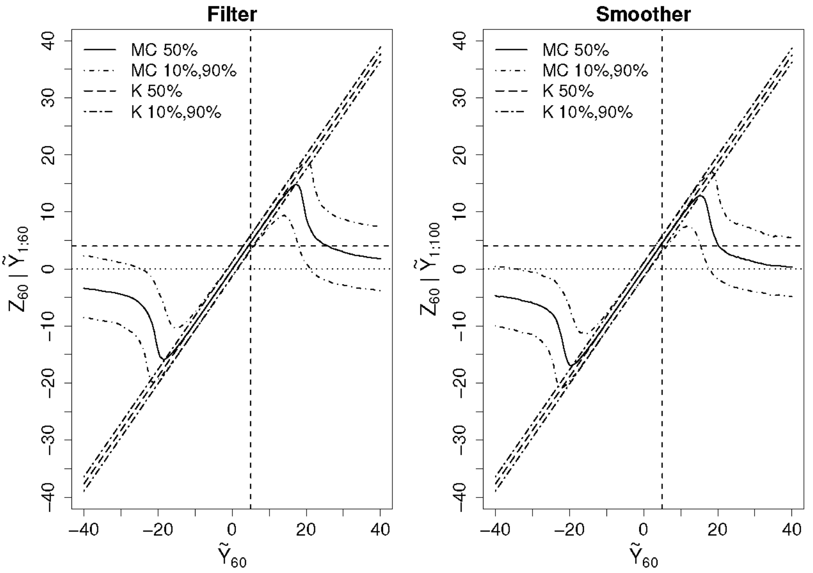

Figure 5.8: *Estimated quantiles using the Kalman and Monte Carlo recursions in the simulation study with one outlier. The time series is simulated with $\varphi_1 = 0.2$, $\sigma_V = 4$ and $\sigma_W = 1$. The Monte Carlo recursions are computed with $m = 3$ and $N = 2000$. The vertical and horizontal dashed lines give the simulated $\widetilde{Y}_{60}$ and Z_{60}, respectively.*

to the simulated state z_{60}. As we can see, the Kalman estimates are not robust against the outlier. Moreover, the width of the $10\% - 90\%$ confidence intervals is always the same since the estimated filtering and smoothing variances are independent of the observations. On the other hand, the Monte Carlo estimates are the same as the Kalman ones for values of $\widetilde{y}_{60}$ around zero. But, with increased $|\widetilde{y}_{60}|$ values, the Monte Carlo recursions detect the outlier in $\widetilde{y}_{60}$, they discredit it and they reduce the estimates. The behaviour is very intuitive, too. First, the quantile nearest zero reacts to the outlier, then the median and, lastly, the other quantile. Furthermore, the width of the $10\% - 90\%$ confidence intervals is not constant but it becomes larger once the outlier is recognized. This reflects the uncertainty about the state estimates.

The values of $\widetilde{y}_{60}$ where the outlier detection takes place depend on the value of the observations around $\widetilde{y}_{60}$. In fact, once the outlier becomes noticeable, it will be discredited. From Figure 5.7, we see that $\widetilde{y}_{59}$ and

$\widetilde{y}_{61}$ are negative ($\widetilde{y}_{61}$ is even -8). This affects the results in Figure 5.8. The outlier is already discovered by $\widetilde{y}_{60} = 16$ for positive values of $\widetilde{y}_{60}$, but only by $\widetilde{y}_{60} = -18$ for negative ones.

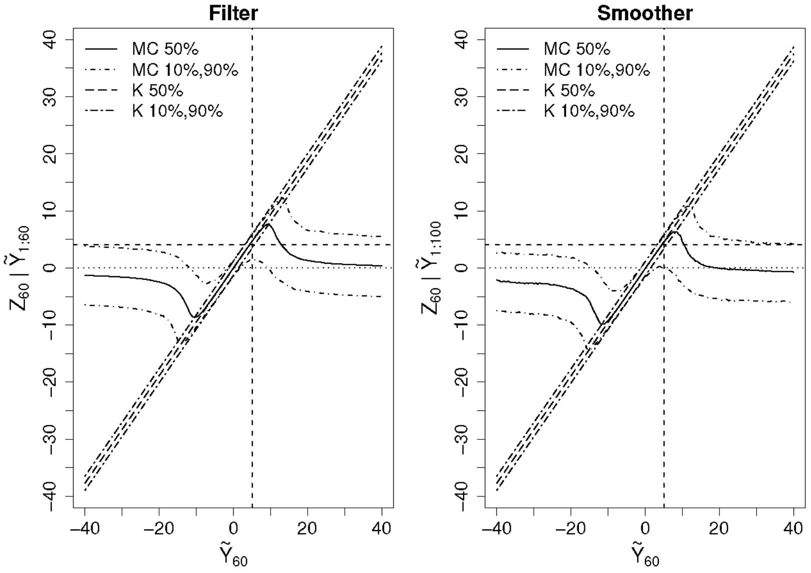

Figure 5.9: *Same as Figure 5.8 with $m = 1$ in the Monte Carlo algorithms.*

The choice of m influences the outlier detection, too. Figure 5.9 shows the quantile estimates for the first time series using $m = 1$ in the Monte Carlo algorithms. As expected, the outlier is found earlier, since the fits with $m = 1$ are more robust.

It is also interesting to analyse what happens to the distributions of the Monte Carlo samples $(z_{60}^{(i)})$ for values of $\widetilde{y}_{60}$ around the outlier detection regions. We illustrate this behaviour using histograms of the samples. For this purpose, we have chosen a quite large sample size N. Thus, Figure 5.10 shows the distributions of the Monte Carlo filtering samples computed with $m = 3$. Looking at the figure from left to right, from top to bottom, we see that the distribution of $(z_{60}^{(i)})$ is unimodal for $\widetilde{y}_{60} = -23.5$ and the main mass is around the value -8. This corresponds to the fact that the observation $\widetilde{y}_{60} = -23.5$ is detected as an outlier. With $\widetilde{y}_{60}$ around -21, the estimates of Z_t are more uncertain. The distribution becomes bimodal with peaks around the values $\widetilde{y}_{60}$ and

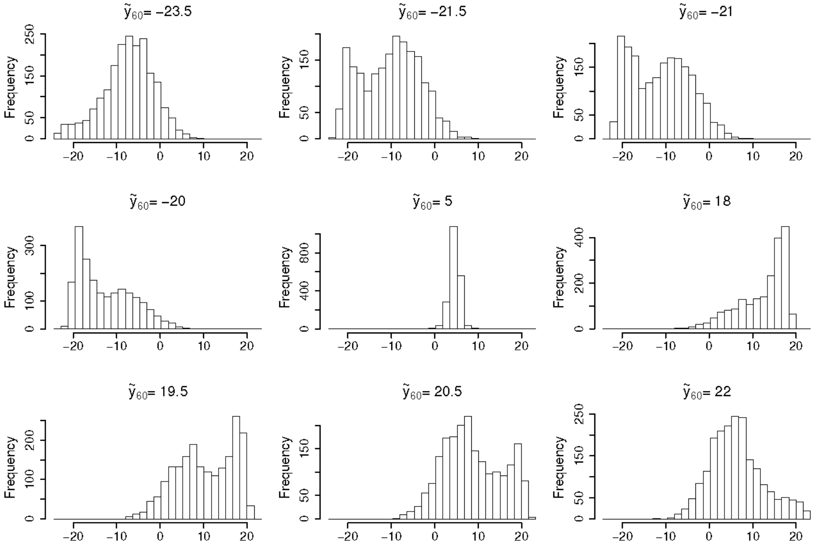

Figure 5.10: *Histograms of the Monte Carlo filtering samples at time* $t = 60$ *for different* $\widetilde{y}_{60}$ *values (simulation study with one outlier). The time series is simulated with* $\varphi_1 = 0.2$, $\sigma_V = 4$ *and* $\sigma_W = 1$. *The Monte Carlo recursion is computed with* $m = 3$ *and* $N = 2000$.

-8, as before. This value of $\widetilde{y}_{60}$ is in the outlier detection region (see Figure 5.8). The observation $\widetilde{y}_{60}$ is not considered an outlier any more. This is reflected by the fact that the peak near $\widetilde{y}_{60}$ is still dominated by the one around -8 for $\widetilde{y}_{60} = -21.5$. But the former one becomes dominating with $\widetilde{y}_{60} = -21$. With $\widetilde{y}_{60} = -20$, there is only the peak near $\widetilde{y}_{60}$. The distribution is again unimodal and the observation $\widetilde{y}_{60}$ is not detected as an outlier any more. Then, the distribution moves from the negative to the positive axis until $\widetilde{y}_{60}$ comes near the outlier detection region on the positive side. Now, it happens the contrary of before. With $\widetilde{y}_{60} = 18$, there is only one peak near $\widetilde{y}_{60}$. With $\widetilde{y}_{60} = 19.5$, there are two peaks near 8 and $\widetilde{y}_{60}$ and the latter one is dominating. With $\widetilde{y}_{60} = 20.5$, there are again the same peaks but the one near 8 is dominating. Finally, with $\widetilde{y}_{60} = 22$, it remains only the peak near 8.

The same behaviour can be seen in the distributions of the Monte Carlo smoothing samples, but at some slightly different values (see Figure 5.11).

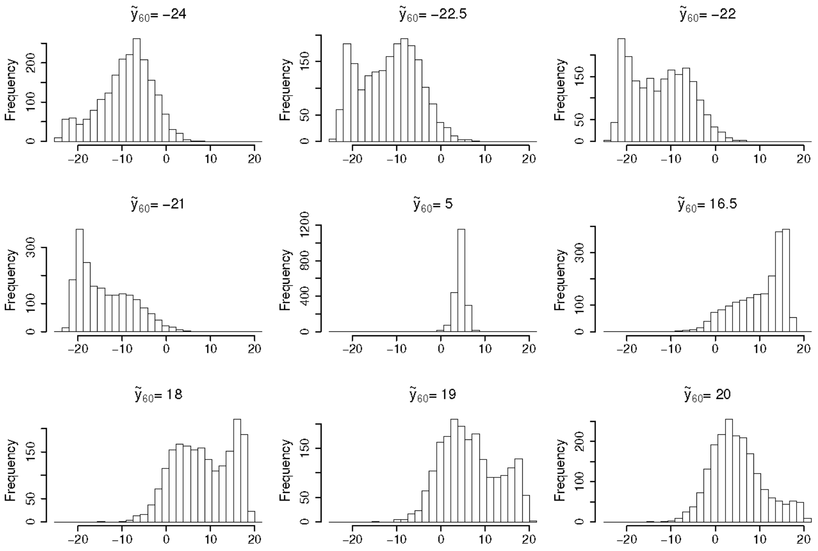

Figure 5.11: *Same as Figure 5.10 using the Monte Carlo smoothing samples.*

Finally, we consider the second time series. It has the same coefficients as one time series studied in Subsection 5.1.1, but different length. A section of it around $t = 60$ is shown in Figure 5.7 and the estimated quantiles of Z_t are shown in Figure 5.12. As we can see, the results are similar to the previous ones. The histograms of the samples present the same behaviour as above, but less clearly.

In addition, we can note in Figures 5.8, 5.9 and 5.12 that the Monte Carlo smoothing algorithm reacts a little bit earlier to the outlier than the filtering one. This is a consequence of the fact that the smoothing algorithm uses the whole set of observations $\widetilde{Y}_{1:100}$.

5.2 Parameter estimation

In this section, we compare the parameter estimates obtained by the developed Monte Carlo algorithms with the Gaussian maximum likelihood estimates computed using the Kalman filtering recursion. The comparison is carried out by a simulation study. We use a Gaussian

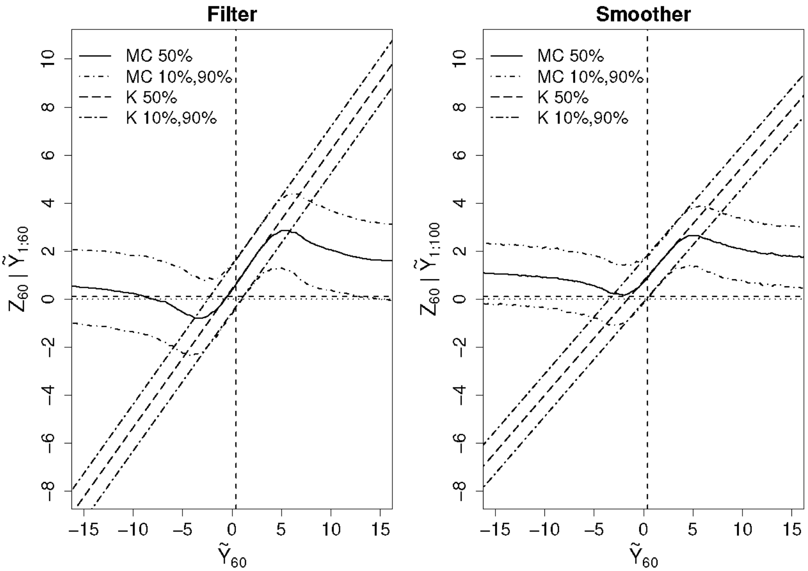

Figure 5.12: *Same as Figure 5.8, with* $\varphi_1 = 0.8$, $\sigma_V = 1$ *and* $\sigma_W = 1$.

state space model and we consider two different cases: without outliers and with 10% isolated or patchy (of length 5) additive outliers. Explicitly, in the first case time series $\widetilde{Y}_{1:T}$ are simulated according to the state space model

$$Z_t = \varphi_1 Z_{t-1} + \cdots + \varphi_p Z_{t-p} + V_t, \qquad V_t \sim \mathcal{N}(0, \sigma_V^2), \tag{5.3}$$

$$\widetilde{Y}_t = Z_t + W_t, \qquad W_t \sim \mathcal{N}(0, \sigma_W^2). \tag{5.4}$$

In the second case, the time series $\widetilde{Y}_{1:T}$ are generated according to

$$Z_t = \varphi_1 Z_{t-1} + \cdots + \varphi_p Z_{t-p} + V_t, \qquad V_t \sim \mathcal{N}(0, \sigma_V^2), \tag{5.5}$$

$$\widetilde{Y}_t = Z_t + W_t + O_t, \qquad W_t \sim \mathcal{N}(0, \sigma_W^2) \tag{5.6}$$

with

$$O_t = \begin{cases} k \cdot sd(w_t) \cdot U_t & \text{if } t \in I_{out}, \\ 0 & \text{otherwise.} \end{cases}$$

We choose a 10% outlier contamination and we denote the outlier index subset by I_{out}. We consider both the situation with isolated outliers and the one where the outliers occur in patches of length 5. In both cases, the outlier indices are chosen randomly. The sequence (U_t), $t \in I_{out}$, gives the outlier signs: U_t takes value in $\{-1, 1\}$ with probability $\{0.5, 0.5\}$. For the patchy case, the values of (U_t) are the same on each patch. Moreover, the outlier magnitude is given by the term $k \cdot sd\,(w_t)$ where $sd(w_t)$ is the empirical standard deviation of the realized observation errors W_t and k is a multiplicative factor. We choose several values of k. Finally, the models (5.3), (5.4) and (5.5), (5.6) are examined using two different state space equations: an AR(1) and an AR(4) processes.

We use the smoothing and the MCEM methods to compute the estimates, see their definitions in Sections 4.1 and 4.2. We do 30 iterations, and the required initial estimates are found using the starting function described in Section 4.3. Since the estimates computed with both Monte Carlo methods fluctuate randomly around the true maximum likelihood value $\widehat{\theta}_{ML}$, we take the median of the last 5 estimates $\widehat{\theta}_{26}, \ldots, \widehat{\theta}_{30}$ to get the final estimate $\widehat{\theta}$ in both methods.
In addition, we compute the estimates using several values for the degrees of freedom m and the sample size N. If we do not fix a value of m a priori, the estimate $\widehat{\theta}$ which maximizes the approximate log-likelihood over the different starting values of m is chosen as final estimate of the unknown coefficients. This approximate log-likelihood is given by

$$\begin{aligned} l(\theta) &= \log\Big[p\left(\{\widetilde{y}_t, t \in I_{av}\}\,|\theta\right)\Big] \\ &= \sum_{t \in I_{av}} \log\Big[p\left(\widetilde{y}_t|\,\{\widetilde{y}_t, t \in \{1, \ldots, t-1\} \cap I_{av}\}\,, \theta\right)\Big] \\ &= \sum_{t \in I_{av}} \log\Bigg[\int p\left(\widetilde{y}_t|z_t, \{\widetilde{y}_t, t \in \{1, \ldots, t-1\} \cap I_{av}\}\,, \theta\right) \cdot \\ &\qquad\qquad \cdot\, p\left(z_t|\,\{\widetilde{y}_t, t \in \{1, \ldots, t-1\} \cap I_{av}\}\,, \theta\right) dz_t\Bigg] \\ &\approx \sum_{t \in I_{av}} \log\left[\frac{1}{N}\sum_{i=1}^{N} p\left(\widetilde{y}_t|z_t^{(i)}, \theta\right)\right] \end{aligned} \tag{5.7}$$

with $(z_t^{(i)})$ a sample of $Z_t|\big\{\widetilde{Y}_t, t \in \{1, \ldots, t-1\} \cap I_{av}\big\}, \theta$. I_{av} is given as in Definition 4.1 and the sample $(z_t^{(i)})$ is produced by simulating from the state equation using the filter sample $(z_{t-p:t-1}^{(i)})$.

On the other hand, the exact log-likelihood function for the Gaussian state space model (1.1), (1.2) can be computed from the Kalman filtering recursion. We have

$$\begin{aligned} l(\theta) &= \sum_{t \in I_{av}} \log \left[p\left(\widetilde{y}_t | \{\widetilde{y}_t, t \in \{1, \ldots, t-1\} \cap I_{av}\}, \theta\right) \right] \\ &= -0.5 \sum_{t \in I_{av}} \left[\log(2\pi) + \log(M_t) + \frac{1}{M_t}\left(\widetilde{y}_t - \mu_{t|t-1}\right)^2 \right] \end{aligned}$$

since $\widetilde{Y}_t | \{\widetilde{Y}_t, t \in \{1, \ldots, t-1\} \cap I_{av}\}, \theta$ is $\mathcal{N}(\mu_{t|t-1}, M_t)$ distributed. We have

$$\begin{aligned} \mu_{t|t-1} &= H_t m_{t|t-1}, \\ M_t &= H_t R_{t|t-1} H_t^{'} + \Omega_t \end{aligned}$$

where $m_{t|t-1}$ and $R_{t|t-1}$ are the Kalman prediction mean and variance, respectively, see (1.3) and (1.4). Note that if $\widetilde{Y}_{t'}$ is missing, we set $m_{t'|t'} = m_{t'|t'-1}$ in the Kalman filtering recursion.
Then, the negative log-likelihood function can be minimized to obtain the maximum likelihood estimate $\widehat{\theta}_{ML}$. The minimization is carried out by the *R* function *optim* which requires starting estimates, too. We take the ones used to begin our Monte Carlo estimation methods, i.e. the ones computed by the starting function described in Section 4.3.
This maximum likelihood method derived from the Kalman filter is used to estimate the parameters in the models (5.3), (5.4) and (5.5), (5.6). Of course, in the first case we have exactly the right situation for the Kalman method (Gaussian errors). It is interesting to analyse what happens to the estimates using the second model.

5.2.1 Estimates using an AR(1) state process

We discuss the results found using an AR(1) state process. The length of the simulated time series is 200 and the coefficients are $\varphi_1 = 0.8$, $\sigma_V = 1$ and $\sigma_W = 1$.

First, 100 different time series are simulated from the Gaussian state space model without outliers, see (5.3), (5.4). The resulting estimates are shown in Figure 5.13. “st” denotes the estimates obtained by the starting algorithm (see Section 4.3); “MC: $\widehat{m} = 1$” to “MC: $\widehat{m} = 5$” the estimates found using the smoothing or the MCEM algorithm. The sample size N is 200. In addition, we select between the estimates “MC: $\widehat{m} = 1$” to “MC: $\widehat{m} = 5$” the one which maximizes the approximate log-likelihood (5.7) for each simulated time series. This estimate is denoted

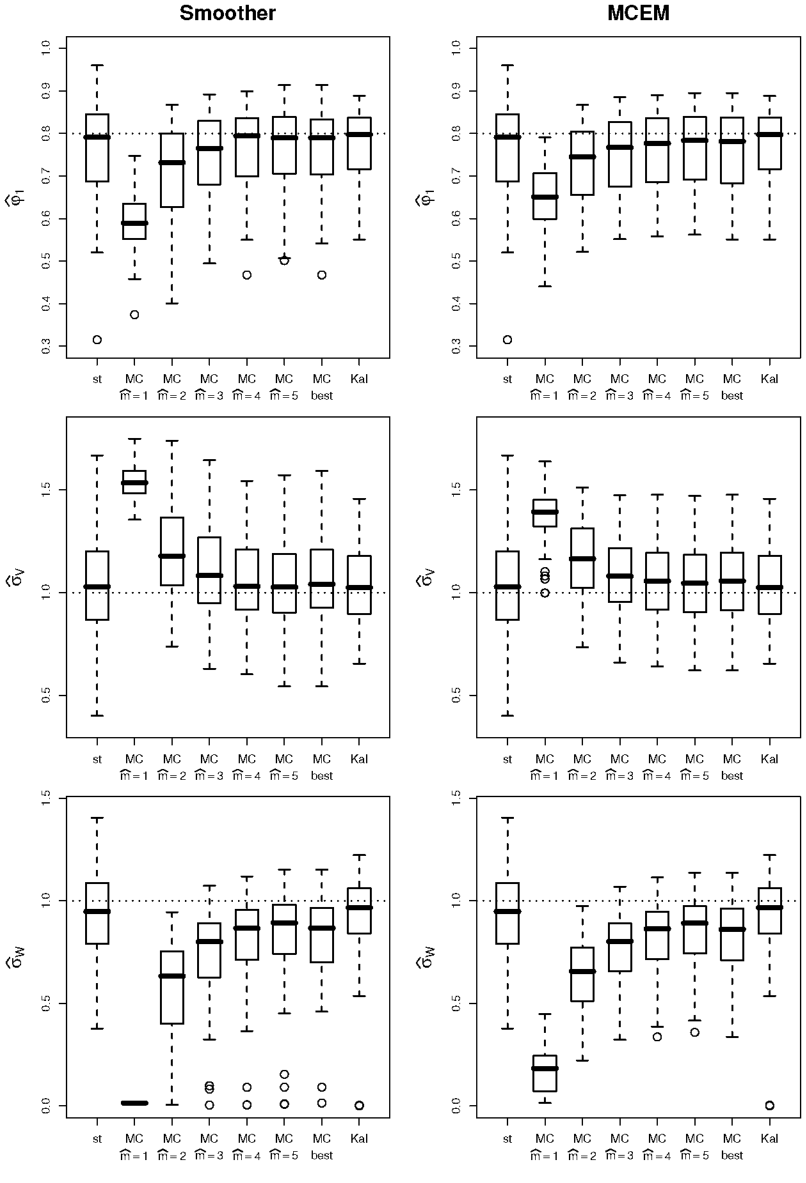

Figure 5.13: *Parameter estimates of the Gaussian AR(1) state space model without outliers. The sample size for the Monte Carlo methods is* 200.

by “MC: best”. Moreover, the maximum likelihood estimates obtained using the Kalman filtering recursion are denoted by “Kal”. The results “st” and “Kal” are shown in both “Smoother” and “MCEM” plots. Finally, for the Monte Carlo smoothing and MCEM algorithms, $\widehat{\sigma}_W$ is found from $\widehat{m}$ and $\widehat{c}$ by comparing the interquartile distance of the normal distribution with the one of the Pearson type VII, see before. We have

$$\widehat{\sigma}_W = \frac{\widehat{c}}{\sqrt{2\widehat{m}-1}} \, \frac{q_{t_{2\widehat{m}-1}}(0.75)}{q_{\mathcal{N}(0,1)}(0.75)}.$$

The true coefficient is given in each plot by the dotted horizontal line. As we can see, the starting estimates are not bad. It is also to be expected that the “MC: $\widehat{m} = 1$” results are not good for the exact Gaussian model. In fact, this method is too robust and it interprets the time series as a AR(1) process with a big σ_W and, consequently, a small φ_1. The “MC: $\widehat{m} = 2$” method has the same problem, to a lesser degree. The other Monte Carlo methods produce very similar estimates and log-likelihood values. The best Monte Carlo estimates are actually found between the estimates computed with these methods, see Table 5.4. The estimates $\widehat{\varphi}_1$ and $\widehat{\sigma}_V$ for “MC: $\widehat{m} = 3$”, ..., “MC: $\widehat{m} = 5$” and “MC: best” are also comparable with the Kalman ones and with the true values. As expected, the dispersion is a little bit greater. On the other hand, the estimates $\widehat{\sigma}_W$ are smaller than both the corresponding Kalman estimates and the true value. This can be a consequence of the small over-estimate of σ_V. We should also not forget that $\widehat{\sigma}_W$ is not found directly but it is derived by comparing the interquartile distances. Smoothing and MCEM estimates are similar, except for $\widehat{m} = 1$. The advantage of the smoothing method is that it converges faster than the MCEM one. But it has bigger fluctuations at the convergence. Figure 5.14 illustrates these features for two time series. The Monte Carlo estimates are computed with $\widehat{m} = 5$. For both the Kalman and

MC	$\widehat{m}=1$	$\widehat{m}=2$	$\widehat{m}=3$	$\widehat{m}=4$	$\widehat{m}=5$
Smoother	2	2	15	32	49
MCEM	0	3	14	37	46

Table 5.4: *Frequencies of the value of $\widehat{m}$ which gives the best Monte Carlo smoothing or MCEM estimate of the parameters of the Gaussian AR(1) state space model without outliers. The Monte Carlo sample size is* 200.

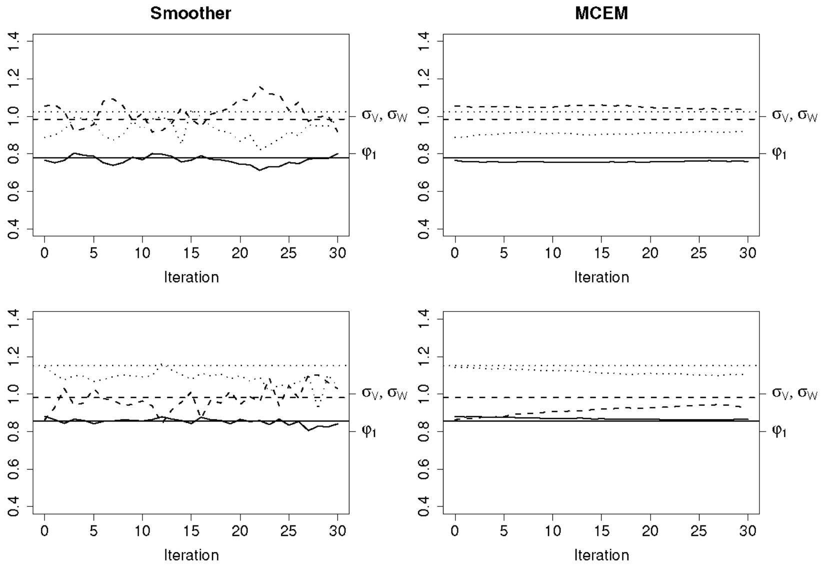

Figure 5.14: *Iterative estimates for two time series (upper and lower panels, respectively) in the Gaussian AR(1) state space model without outliers. The Monte Carlo estimates are computed with $\widehat{m} = 5$, the sample size N is* 200.

the Monte Carlo methods, the estimates $\widehat{\varphi}_1$, $\widehat{\sigma}_V$ and $\widehat{\sigma}_W$ are shown by solid, dashed and dotted lines, respectively. The Kalman estimates are the horizontal lines in the plots. The true parameters are given on the right margin.

It is also interesting to examine what happens if we take a different sample size N, for example 500. The results are shown in Figures 5.15 and 5.16. The considered time series are the same as before for both figures. As we can see, the estimates are very similar. In these examples, a greater value of N does not reduce the fluctuations of the Monte Carlo smoothing estimates.

The next point to discuss is the influence of isolated or patchy outliers on the estimates. This is analysed by simulating an AR process (Z_t) and a Gaussian white noise sequence (W_t) as described in (5.5) and (5.6), respectively. Then, the outlier index subset I_{out} and the signs (U_t), $t \in I_{out}$, are generated. Finally, the observations $(\widetilde{Y}_t)$ are calcu-

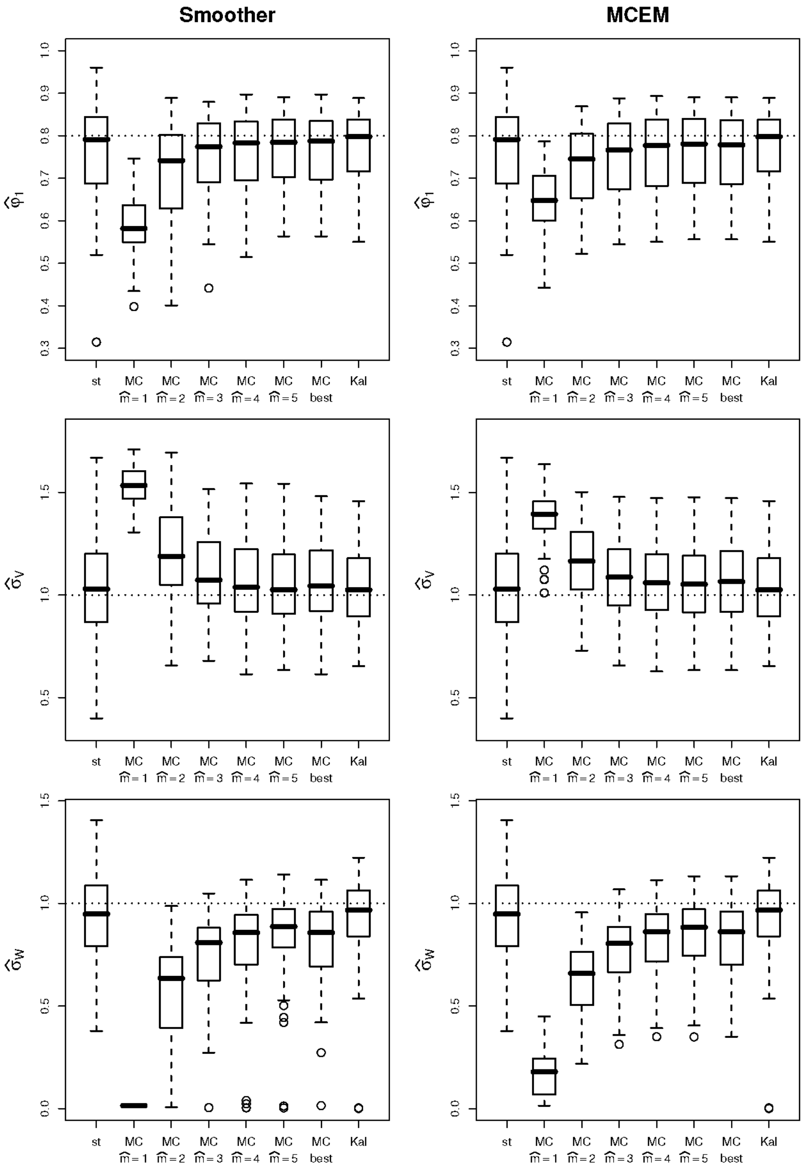

Figure 5.15: *Same as Figure 5.13, with $N = 500$.*

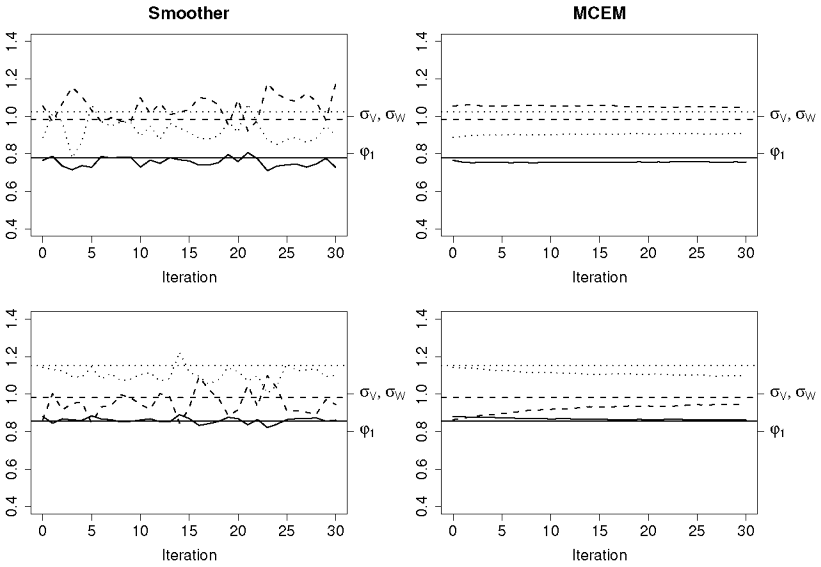

Figure 5.16: *Same as Figure 5.14, with* $N = 500$.

lated using several values of the magnitude k of the outliers, see (5.6), and the estimates are computed depending on the values of k. For the AR(1) state equation, we choose k in $\{\text{NA}, 0, 2, 3, 5, 10, 15, 20\}$. With k equal to NA, we denote the case where the observations $\{\widetilde{Y}_t, t \in I_{out}\}$ are not available.

Figures 5.17 and 5.18 show the results for two different sequences (Z_t) and (W_t). Note that we show only two examples, since the behaviour is similar for all considered sequences. In both figures, the starting estimates are connected by the black lines; the Monte Carlo MCEM with $\widehat{m}$ in $\{1, 2, 3\}$ using the green solid, dashed, and dotted lines, respectively. The Monte Carlo estimates with $\widehat{m} = 4$ and $\widehat{m} = 5$ are omitted for the sake of clarity. The Monte Carlo estimates which maximize the approximate log-likelihood (5.7) with respect to $\widehat{m}$ in $\{1, 2, 3, 4, 5\}$ are connected by the blue lines. Finally, the Kalman estimates are connected by the red lines.

We can see that all estimates are good for small values of k except the Monte Carlo ones with $\widehat{m} = 1$. As before, this is to be expected since the outliers are still tolerable. The starting estimates are not bad either. With increasing magnitude k, the Kalman estimates and sometimes the

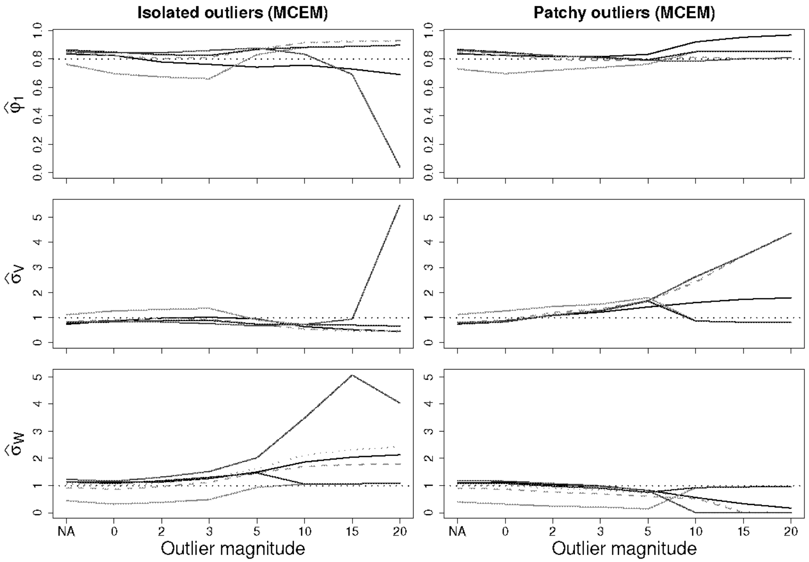

Figure 5.17: *First example of the influence of* 10% *outliers on the estimates of the parameters of the AR(1) state space model. The Monte Carlo sample size N is* 200. *See text for the other notations.*

starting estimates become inaccurate. On the other hand, the Monte Carlo method with $\widehat{m} = 1$ produces reliable results and the best Monte Carlo estimates are given by it. The other Monte Carlo estimates become less accurate with increasing $\widehat{m}$ and large k (including the omitted estimates with $\widehat{m} = 4$ and $\widehat{m} = 5$). In addition, the estimates in the case where 10% of the observations are not available are similar to the ones without outliers.

We note that the Kalman estimates of φ_1 also remain good with big magnitudes for patchy outliers. This is a consequence of the fact that the signs (U_t) are the same on each patch. Therefore, the AR structure is not disturbed too much, only the scale parameters σ_V and σ_W are affected by the outliers.

Figure 5.18 shows that the resulting Monte Carlo estimates are also reliable with bad starting values.

As before, we can examine what happens with a different Monte Carlo sample size N, for example 500. Figure 5.19 shows the results for the same sequences (Z_t) and (W_t) used to get Figure 5.18. The estimates

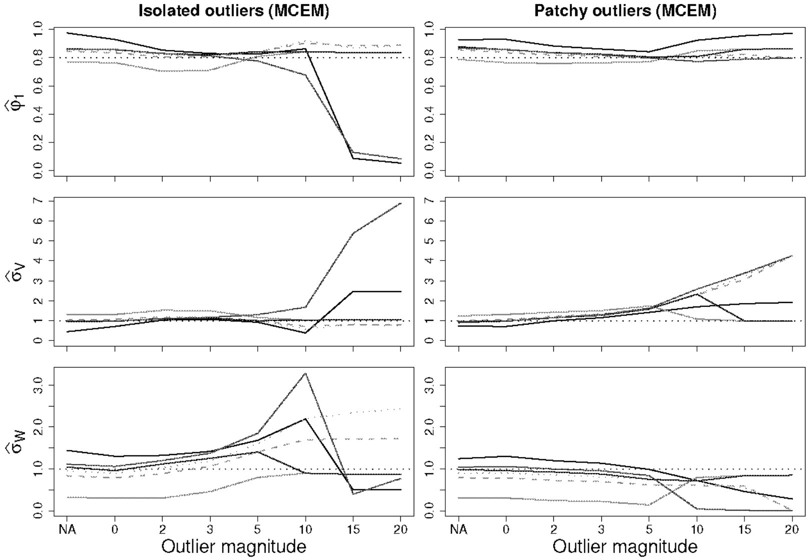

Figure 5.18: *Second example of the influence of* 10% *outliers on the estimates of the parameters of the AR(1) state space model. The Monte Carlo sample size* N *is* 200. *See text for the other notations.*

are practically the same.

The estimates obtained by the Monte Carlo smoothing method present the same features as the MCEM ones. For this reason, they are not shown.

5.2.2 Estimates using an AR(4) state process

In the second case, we choose a more difficult AR process as state equation. The AR coefficients are as in Percival and Walden (1993), Section 2.6.: $\varphi_1 = 2.7607$, $\varphi_2 = -3.8106$, $\varphi_3 = 2.6535$, $\varphi_4 = -0.9238$. This is an interesting example to test the developed algorithms since this AR process is near the non-stationarity border and the (logarithmic) spectral density is bimodal, see for example Figure 5.23. The scale parameters are $\sigma_V = 1$ and $\sigma_W = 1$ and the length of the simulated time series is 256. As before, we analyse the resulting parameter estimates of the models (5.3), (5.4) and (5.5), (5.6). In addition, we consider the

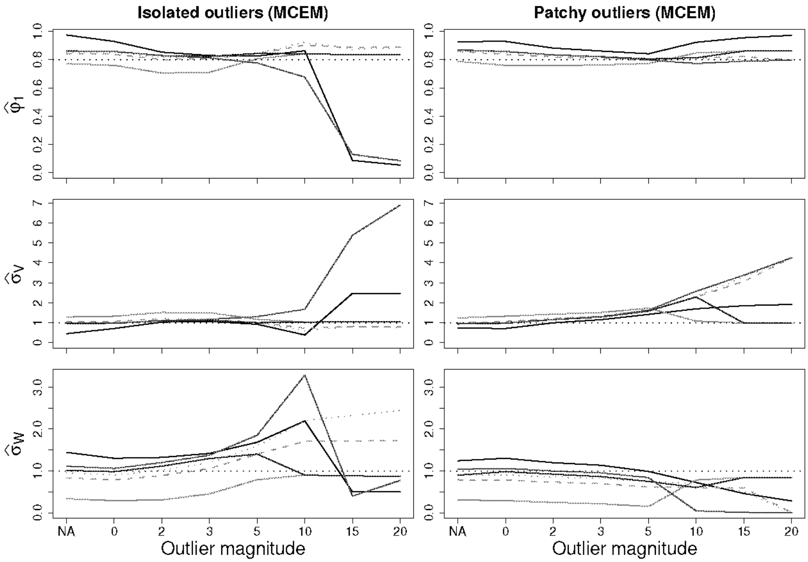

Figure 5.19: *Same as Figure 5.18, with $N = 500$.*

estimated (logarithmic) spectral densities for the second model. It is very illustrative in this example.

First, 50 different time series are simulated according to the Gaussian model without outliers, see (5.3) and (5.4). The resulting estimates are shown in Figure 5.20. The Monte Carlo estimates are computed with the MCEM method and the notation is the same as before.
As we can see, the starting estimates are no more accurate. This is somewhat surprising, but we should not forget that the chosen AR process is near the non-stationarity border. Nevertheless, the Kalman estimates are good. The MCEM estimates are quite good for small values of $\widehat{m}$ and "MC: best". An exception are the estimates of σ_W. A reason could be as before that the estimates are computed by comparing the interquartile distances and not directly.
The Monte Carlo estimates computed with the smoothing method are very similar to the MCEM ones. There are some more outliers and some box plots are a little bit wider. The estimates $\widehat{\sigma}_W$ are better for "MC: $\widehat{m} = 4$" and "MC: $\widehat{m} = 5$" (the 50% box contains the true value).

The influence of isolated or patchy (of length 5) outliers is examined

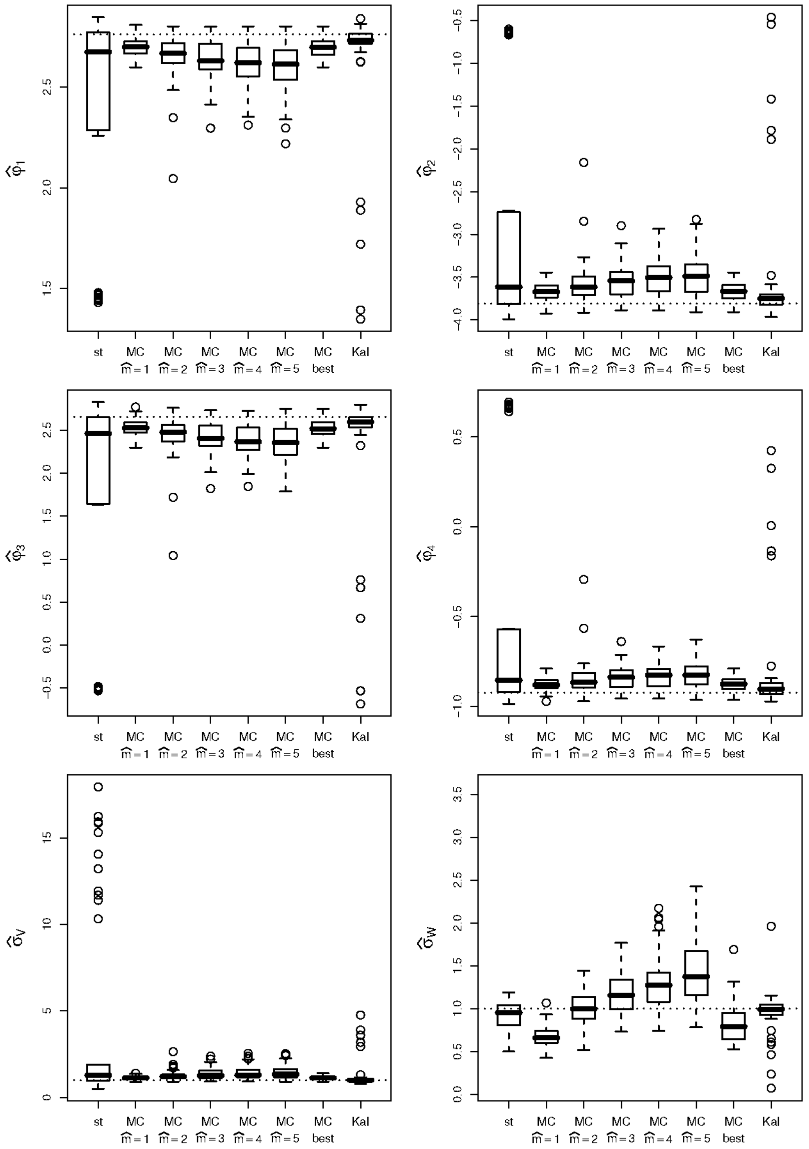

Figure 5.20: *Parameter estimates of the Gaussian AR(4) state space model without outliers. The sample size for the Monte Carlo MCEM methods is* 200.

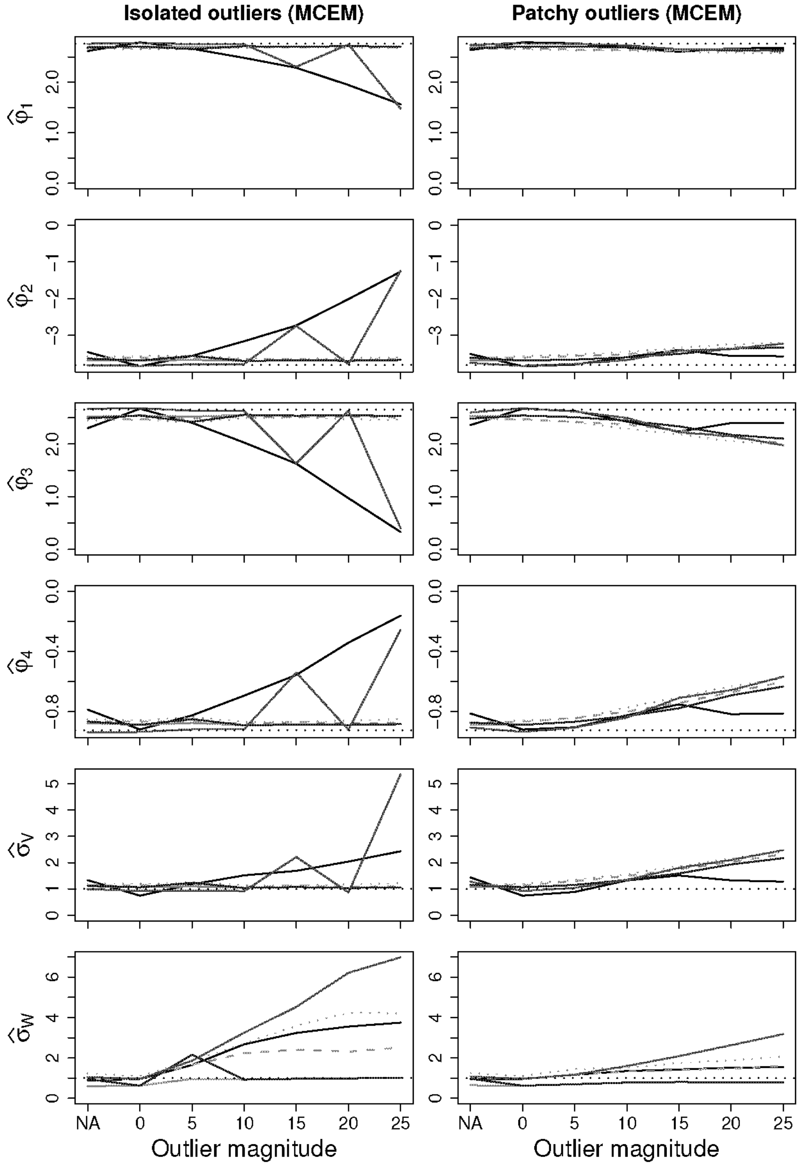

Figure 5.21: *First example of the influence of* 10% *outliers on the estimates of the parameters of the AR(4) state space model. The Monte Carlo MCEM sample size N is* 200. *See text for the other notations.*

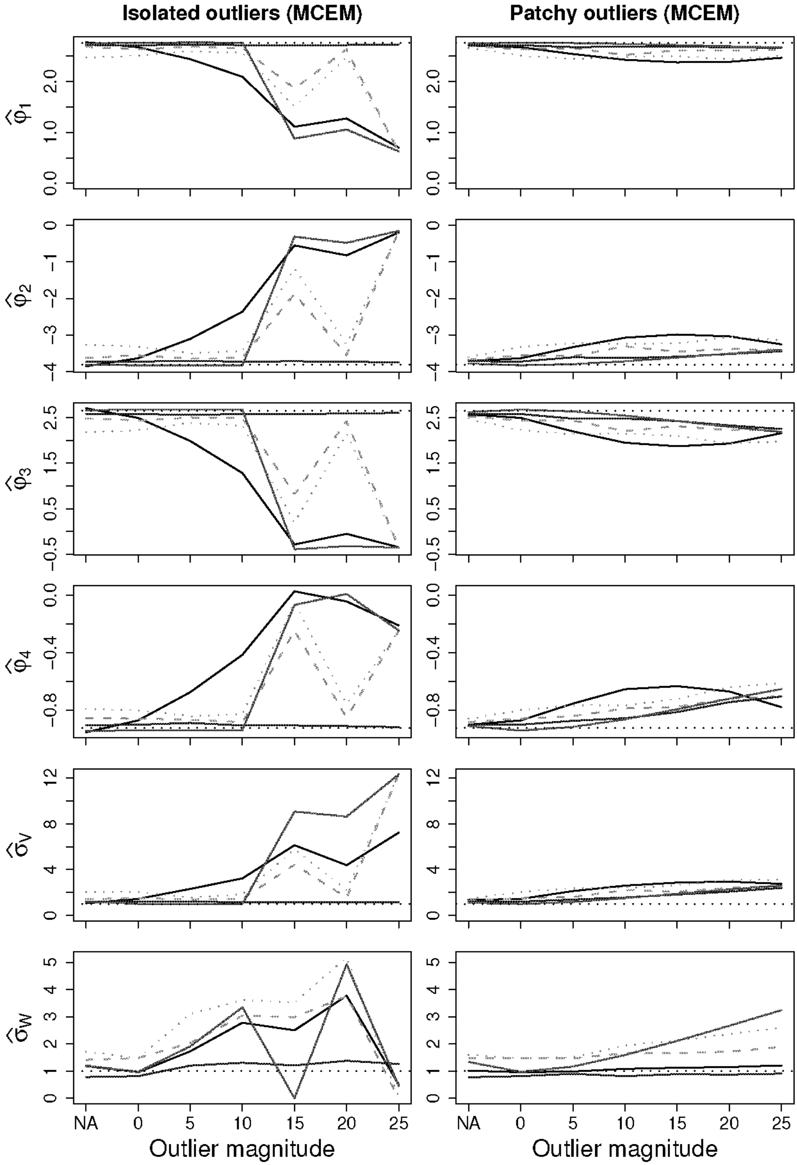

Figure 5.22: *Second example of the influence of* 10% *outliers on the estimates of the parameters of the AR(4) state space model. The Monte Carlo MCEM sample size N is* 200. *See text for the other notations.*

using the model (5.5) and (5.6). The outlier magnitude k takes values in the set $\{\mathrm{NA}, 0, 5, 10, 15, 20, 25\}$. This set is adapted to the fact that the observations have larger values with this AR(4) state process. The results for two sequences (Z_t) and (W_t) are shown in Figures 5.21 and 5.22. Note that we show only two examples, since the behaviour is similar for all considered sequences. In addition, the Monte Carlo estimates are computed again with the MCEM method. The colors and the line types have the same meaning as before.

In the case with isolated outliers, the Monte Carlo estimates with $\widehat{m} = 1$ (and then the best Monte Carlo estimates) remain accurate for all outlier magnitudes k. This is not true for the starting estimates and the Kalman ones. Their estimates become worse with increasing k. Unfortunately, the estimates with patchy outliers are not so good. In fact, the Monte Carlo estimates become imprecise with increasing k, This is due to the fact that the considered model (2.1), (2.2) presupposes independent observation errors.

We can illustrate these results better considering the logarithmic spectral density of the AR(4) state process. It is given by

$$\log(f(\nu)) = 2\log(\sigma_V) - \log\left(\left|1 - \sum_{l=1}^{4} \varphi_l \exp(-i2\pi\nu l)\right|^2\right).$$

It is bimodal for the chosen AR(4) process.

Figures 5.23 and 5.24 show the results for the same two sequences (Z_t) and (W_t) of Figures 5.21 and 5.22. The logarithmic spectral densities computed with the best MCEM, the Kalman and the true coefficients are shown by solid, dashed and dotted lines, respectively. In the case with isolated outliers, we can reproduce well the bimodality for all k using the best MCEM estimates. This is not the case with the Kalman estimates. But also the best MCEM method is no more able to reproduce the two peaks with big values of k in the patchy case.

The plots using the Monte Carlo smoothing estimates are very similar to Figures 5.21, 5.22, 5.23 and 5.24.

5.3 Estimation of road traffic emissions

We analyse now an example with real data. It actually motivated the present thesis.

The emissions of road vehicles contribute significantly to air pollution.

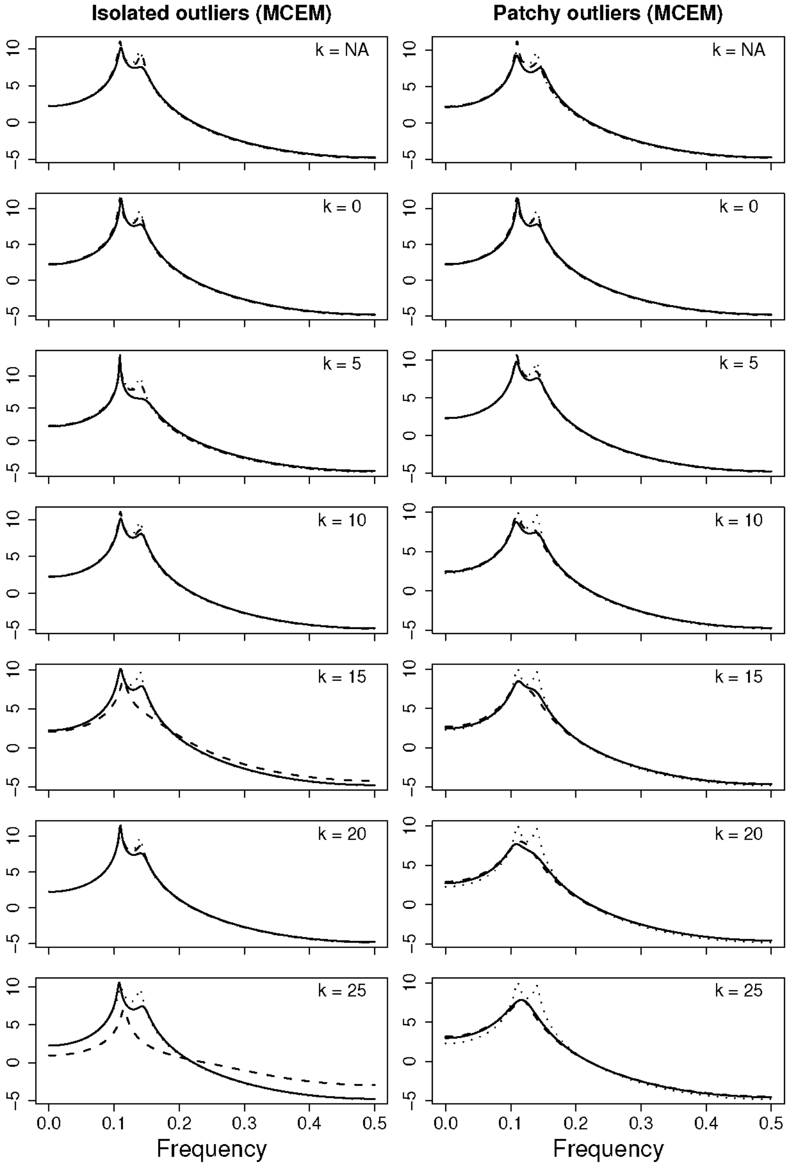

Figure 5.23: *First example of the logarithmic spectral densities for the AR(4) state space model with* 10% *outliers: best MCEM (solid line), Kalman (dashed line) and true (dotted line). The MCEM sample size N is* 200.

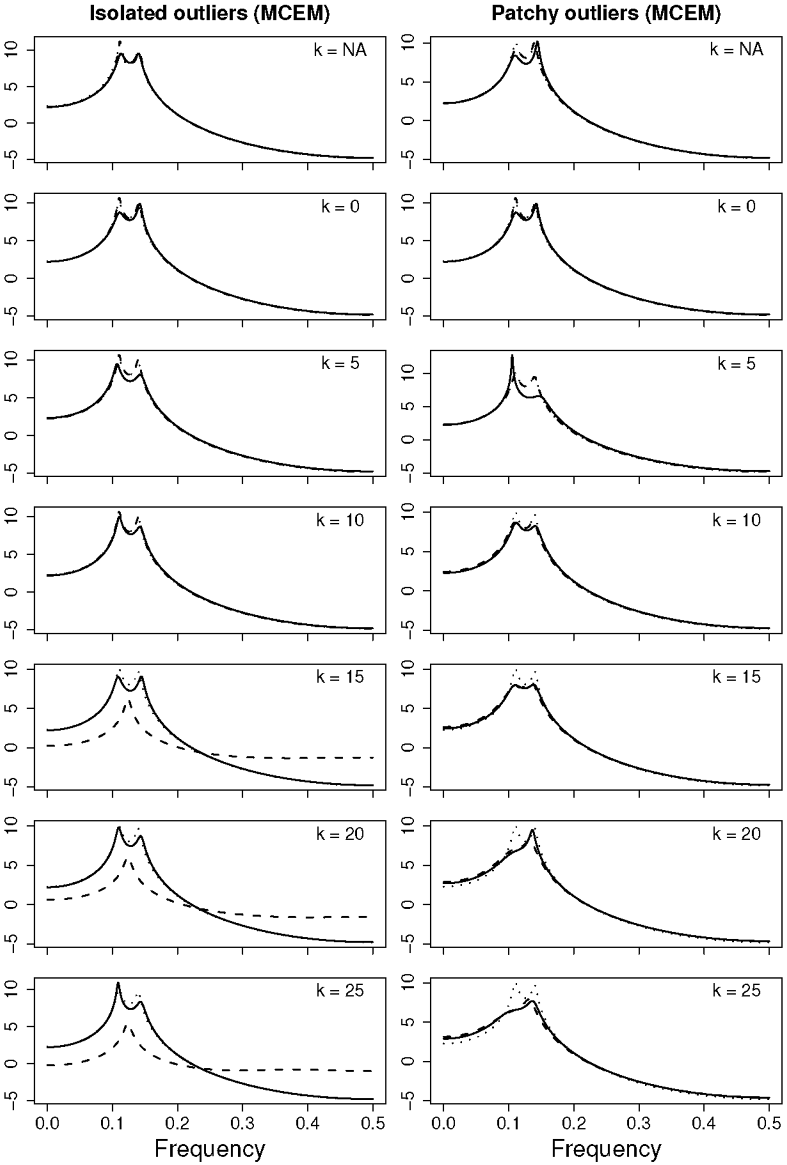

Figure 5.24: *Second example of the logarithmic spectral densities for the AR(4) state space model with* 10% *outliers: best MCEM (solid line), Kalman (dashed line) and true (dotted line). The MCEM sample size* N *is* 200.

A way to characterize the traffic emissions is given by the so-called emission factors (EFs). They give the amount of emitted compounds per driven distance, and they are estimated for single, entire fleet or categories of vehicles. Many variables influence the emission factors, for example size, type, age, cylinder capacity, fuel mode (gasoline or diesel) of the vehicle, presence of the catalytic converter. Or the driving style, the kind of road (highway, ...) and its gradient. Consequently, the estimation of emission factors is quite complex. At least two methods are used to estimate them: dynamometric tests or measurements in tunnels. Dynamometric tests permit to measure the emission factors for single vehicles under given conditions (driving style, speed, road gradient and so on). If these conditions are chosen close to real ones and the tests are carried out on a representative sample of vehicles, then road traffic emission models can be developed. For example, the handbooks of emission factors HBEFA (1999) and HBEFA (2004) describe road traffic emissions of several compounds (CO, NOx, THC) for Austria, Germany and Switzerland. But, unfortunately, the reproducibility of real conditions is not easy.

On the other hand, measurements in a tunnel permit to estimate emission factors in a real traffic situation, although only for some categories of vehicles. Briefly, the measurements are organized as follows. The key idea is that the pollutants are transported with very little losses through a passively ventilated tunnel. Thus, the concentrations at the beginning and at the end of the tunnel are measured for each substance at regular intervals t during one or more weeks, and the concentration differences Δc_t (between the end and the beginning of the tunnel) are computed. Interval shifts are used to compute the differences to take into account the times that the air needs to pass from the beginning to the end of the tunnel. Then, for each pollutant and interval t, the average emitted quantity EF_t per vehicle and driven distance can be computed by multiplying Δc_t by the volume V_t of the air passing through the tunnel in the interval t and dividing by the length d of the tunnel and the number of vehicles n_t passing through it in the interval t. Explicitly,

$$EF_t = \frac{\Delta c_t \cdot V_t}{n_t \cdot d} \tag{5.8}$$

where the volume V_t is obtained by

$$V_t = w_t \cdot s \cdot \Delta t.$$

w_t is the air speed measured inside the tunnel (it depends on the traffic flow), s is the tunnel cross section and Δt is the length of the considered

time interval. Emission factors are usually expressed in g/km. In addition, loop detectors set on the road provide traffic informations for the same intervals. Typically, they register the number of vehicles divided in different categories (the classification is done according to the wheel distance) and their speeds. Thus, we can model the average emissions EF_t using the percentages of vehicles in each class and their speeds. In this way, we can find an estimate of the emission factors for each vehicle category.
Both explained methods (dynamometric tests and tunnel measurements) present advantages and disadvantages as we have seen. The aim is to use the estimated emission factors from tunnel measurements to validate the road traffic emission models developed from dynamometric tests.

In this example, we analyse the measurements performed from Monday 9^{th} September to Friday 13^{th} September 2002 in one tube of the Gubrist tunnel close to Zurich, Switzerland. The Gubrist tunnel is a highway tunnel and the considered tube has two lanes with a road gradient of 1.3%. The considered pollutant is nitrogen oxide (NOx: NO + NO_2). Moreover, the loop detectors divide the traffic in two categories: the heavy duty vehicles (HDV) and the light duty vehicles (LDV), i.e. private cars or vans. Most heavy duty vehicles have a diesel engine, the light duty vehicles have mostly a gasoline engine and a small part a diesel one, especially vans. The time intervals of the measurements have length 1 minute. But we consider averages over 15 minutes.
The Gubrist study is actually more comprehensive. In fact, the measurements go on 5 weeks (from 9^{th} September to 10^{th} October) and carbon monoxide (CO) and total Volatile Organic Compounds (t-VOC) are also measured. We consider a reduced example for illustrative purposes.
The average emission factors computed according to (5.8), the percentages of HDV and the vehicle speeds are shown in Figure 5.25. LDV and HDV have nearly the same speeds and thus we have only one time series of speeds. In addition, we have set the values between 8 p.m. and 6 a.m. to NA since few vehicles are on the road during these hours. Then, the conditions to compute the average emission factors EF_t are not ideal (for example inhomogeneous traffic flows and air speeds w_t). The vertical dotted lines in the plots delimit the different days. The relation between average emission factors EF_t and percentages of HDV is immediately visible.

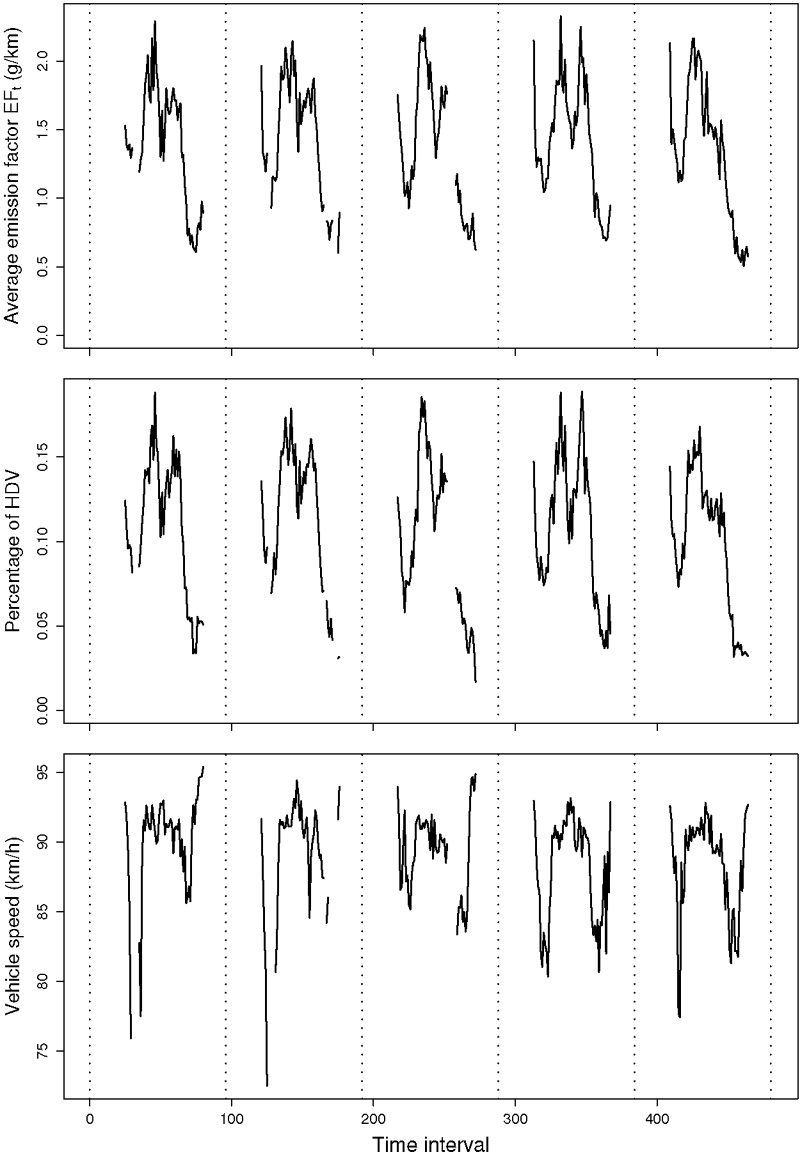

Figure 5.25: *Average emission factors EF_t, percentages of HDV and vehicle speeds for the considered 5 days.*

In previous tunnel studies, a very promising model was given by

$$log(EF_t) = \log(\alpha_1 + \alpha_2 \cdot pHDV_t) + \alpha_3 \cdot speed_t + \varepsilon_t, \tag{5.9}$$

see for example Colberg et al. (2005). In the last model, $pHDV_t$ is the percentage of HDV and $speed_t$ is the vehicle speed on the t^{th} interval. Note that the previous model (5.9) gives

$$EF_t = (\alpha_1 + \alpha_2 \cdot pHDV_t) \cdot \exp(\alpha_3 \cdot speed_t) \cdot \exp(\varepsilon_t),$$

i.e. we have a multiplicative error and the speed scales the emission factor.

Now, the goal is to estimate the parameters α_1, α_2 and α_3. In this way, we can find the estimates of the emission factors of LDV and HDV for different speeds and also their pointwise confidence intervals. To this end, we use different methods. A first estimation is carried out assuming that the errors (ε_t) are independent and Gaussian distributed. We use the *R* function *nls* to compute the nonlinear fit. But, as we could expect, the residuals are correlated and only approximatively normally distributed. Some outliers are also present. Thus, we have the situation covered by our model. We estimate the parameters in (5.9) using the developed MCEM algorithm and the maximum likelihood method derived from the Kalman filtering recursion. We use the starting function described in Section 4.3 to begin both estimation methods. Moreover, we choose an AR(2) process for the errors. This order results considering the autocorrelation and partial autocorrelation plots computed with the residuals of the preliminary nonlinear fit. The order will be confirmed by the MCEM fit. In addition, different estimates of the degrees of freedom m are used to compute the MCEM estimates. The estimate which maximizes the approximate log-likelihood (5.7) over the different values of $\widehat{m}$ is chosen as the final estimate. The number of iterations in the MCEM algorithm is 30 and the sample size is 500.

Once the parameters are estimated, the estimated emission factor $\widehat{EF}$ for given values of $pHDV$ and $speed$ is given by

$$\widehat{EF} = (\widehat{\alpha}_1 + \widehat{\alpha}_2 \cdot pHDV) \cdot \exp(\widehat{\alpha}_3 \cdot speed). \tag{5.10}$$

Its pointwise confidence interval follows from the asymptotic theory of the maximum likelihood estimator. In fact, the maximum likelihood estimate $\widehat{\theta}$ is asymptotically $\mathcal{N}(\theta_0, I(\theta_0)^{-1}/T)$ distributed, see for example Bickel et al. (1998) or Jensen and Petersen (1999). θ_0 denotes the true unknown parameters, $I(\theta_0)$ is the Fisher information and T is the length of the time series. Thus, the estimated emission factor $\widehat{EF}$ is approximatively normally distributed around the true unknown value. The variance is given by

$$\begin{aligned}
Var\left(\widehat{EF}\right) &= Var\left((\widehat{\alpha}_1 + \widehat{\alpha}_2 \cdot pHDV) \cdot \exp(\widehat{\alpha}_3 \cdot speed)\right) \\
&\approx \exp(2\alpha_3 \cdot speed) \cdot Var\left((\widehat{\alpha}_1 + \widehat{\alpha}_2 \cdot pHDV) \cdot \exp(\Delta\alpha_3 \cdot speed)\right) \\
&\approx \exp(2\alpha_3 \cdot speed) \cdot Var\left((\widehat{\alpha}_1 + \widehat{\alpha}_2 \cdot pHDV) \cdot (1 + \Delta\alpha_3 \cdot speed)\right) \\
&\approx \exp(2\alpha_3 \cdot speed) \cdot \\
&\quad \cdot Var\left(\Delta\alpha_1 + \Delta\alpha_2 \cdot pHDV + (\alpha_1 + \alpha_2 \cdot pHDV)\Delta\alpha_3 \cdot speed\right) \\
&\approx \exp(2\alpha_3 \cdot speed) \cdot a^{'} Var\left(\widehat{\theta}\right) a \qquad (5.11) \\
&\approx \frac{1}{T} \exp(2\alpha_3 \cdot speed) \cdot a^{'} I(\theta_0)^{-1} a \qquad (5.12)
\end{aligned}$$

with $\theta_0 := (\alpha_1, \alpha_2, \alpha_3)^{'}$, $a := (1, pHDV, (\alpha_1 + \alpha_2 \cdot pHDV) \cdot speed)^{'}$. Note that the variance of $\widehat{\theta} := (\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3)$ is given by $I(\theta_0)^{-1}/T$.
Now, we use the results (5.10) and (5.12) to estimate the emission factors of LDV and HDV for given speed values and find their approximate pointwise 95% confidence intervals. This is done for the three estimation methods mentioned above (*nls* method, MCEM method and maximum likelihood method derived from the Kalman filtering recursion). We just set $pHDV$ equal to zero to find the estimated emission factors of LDV and equal to one for the HDV. To compute the variances, we substitute the true unknown parameters by the estimated ones. In addition, the Fisher information can be approximated by the Hessian of the negative log-likelihood function computed at the estimated parameters for both the MCEM method and the maximum likelihood method derived from the Kalman filter. For the *nls* method, the variance of $\widehat{\theta}$ in (5.11) can be computed directly from the design matrix of the nonlinear fit, see also below.
It is also interesting to estimate the emission factors of LDV and HDV by applying the method in Colberg et al. (2005). This method is a refinement of the method in Cochrane and Orcutt (1949), see Staehelin et al. (1995). First, the estimate $\widehat{\theta} := (\widehat{\alpha}_1, \widehat{\alpha}_2, \widehat{\alpha}_3)$ is found by fitting a nonlinear model with the function *nls*, see also above. Explicitly, we have

$$\widehat{\theta} = \left(X^{'}X\right)^{-1} X^{'}Y$$

with Y the vector containing all values $\log(EF_t)$ and X the design matrix which is given by the derivatives of the right hand side of (5.9) with respect to the parameters. Then, the residuals are modeled with a robustified AR(2) process and the covariance matrix $\widehat{\Sigma}$ of the residuals is computed. It follows

$$Var(\widehat{\theta}) = \left(X'X\right)^{-1} X'\widehat{\Sigma}X \left(X'X\right)^{-1} / B^2 \tag{5.13}$$

with B a correction term due to the robustification. Then, the estimated emission factors of LDV and HDV are computed using again (5.10), and the confidence intervals are computed using (5.11). The variance in (5.13) is used to compute (5.11).
More details on this method can be found in Staehelin et al. (1995).

The estimated emission factors of LDV and HDV, and their pointwise 95% confidence intervals are shown in Figure 5.26. They are computed with the *nls* method, the refinement of Cochrane and Orcutt (1949) (*nls+AR*) and the MCEM method. The best MCEM estimate is obtained by the degrees of freedom $\widehat{m} = 8$. This means that the observation errors are approximately normally distributed.
Unfortunately, we did not succeed in computing the corresponding results with the maximum likelihood method derived from the Kalman filter. In fact, the optimization routine *optim* of *R* ended always in local minima for the tried starting values (we also used the best MCEM estimate) and the chosen tuning variables (for example the scaling values for the parameters during the optimization). Consequently, we got unreasonable parameter estimates (for example negative values of $\widehat{\alpha}_1$) or nonpositive definite variance matrices $Var\left(\widehat{\theta}\right)$.
In addition, the estimated emission factors from the handbooks HBEFA (1999) and HBEFA (2004) are also shown. The black circles denote the results using the HBEFA (1999), the squares using the HBEFA (2004). These estimates are available only for few speeds. The LDV estimates are computed assuming a mix of 88% cars and 12% vans. This is the average mix in the Gubrist tunnel.
Of course, the emission factors estimated with the *nls* method and the *nls+AR* one are identical since they are found with the same method. The pointwise 95% confidence intervals computed with the *nls+AR* method are wider since the (positive) error correlation is considered.
As expected, the speed has a positive effect on the emission factors of LDV and HDV for all three methods and the two handbooks. But the MCEM estimated emissions of LDV are greater than the *nls* or *nls+AR* ones. The contrary is true for HDV.
Which results are reliable? A well known problem affecting the HBEFA

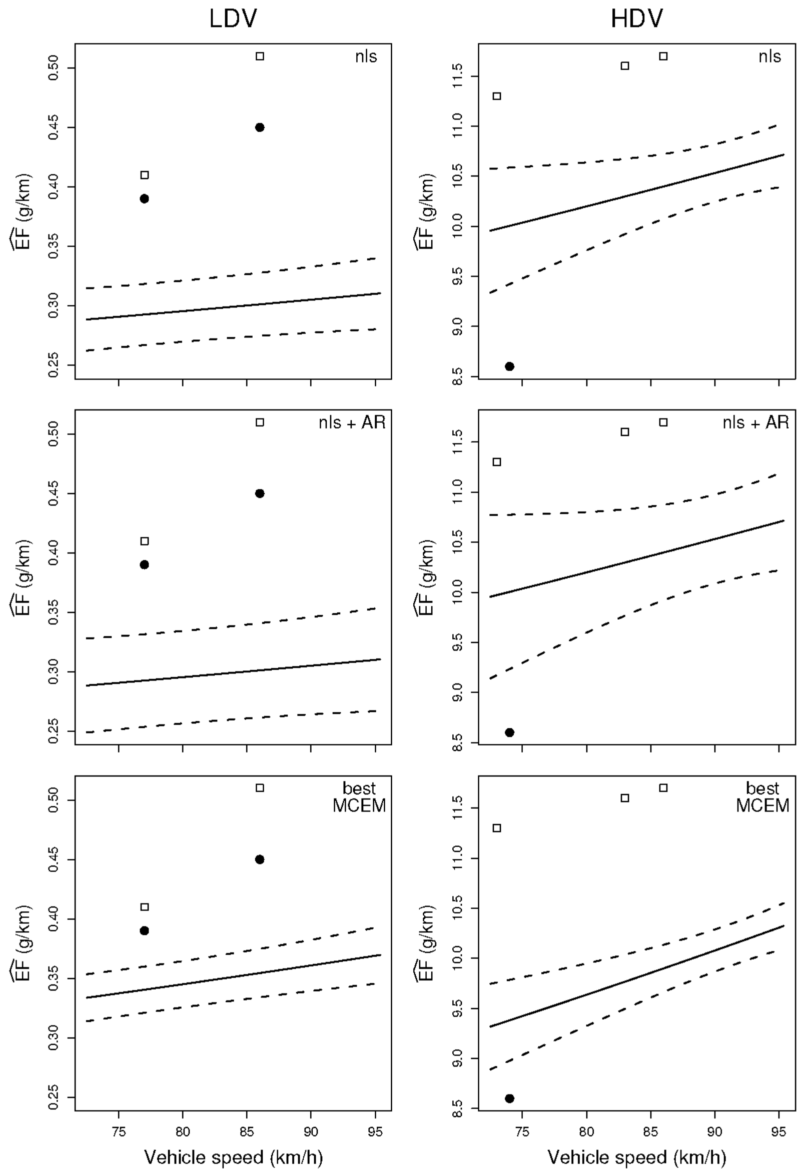

Figure 5.26: *Estimated emission factors and their pointwise* 95% *confidence intervals using the* nls, nls+AR *and best MCEM methods. The black circles denote the results using the HBEFA 99, the squares using the HBEFA 04.*

(1999) is that the emission factors of HDV are systematically underestimated. A reason is that modern diesel engines are equipped with electronic systems which can control and reduce the emissions in steady state approval tests, see for example Hausberger et al. (2002). (Dynamometric tests are performed under the same conditions.) But outside these test conditions, the emissions are greater. This result is confirmed by the three statistical methods. The HBEFA (2004) corrects this problem, but maybe it overestimates now the emissions. Or other reasons to explain the discrepancy between the statistical results and the HBEFA (2004) may be the following ones. The estimates for HDV are computed by setting $pHDV$ equal to one. This extrapolation could be problematic since the observed percentages of HDV never exceed 18%, see Figure 5.25. Or it could be that the vehicle fleet passing through the Gubrist tunnel is "cleaner" than the one assumed in the handbook. In addition, the handbook requires the specification of the mix between cars and vans to compute the estimates for LDV. The mix may also vary considerably during the day. This fact can affect the results since vans have typically a diesel engine. On the other hand, the statistical methods have the disadvantage that the classifications in the two categories LDV and HDV are done using loop detectors. Then, misclassifications are possible. This could be checked by comparing the classifications from the loop detectors with the ones from video records, if the latter are available.

The issue of finding good road traffic emission models is still open. It also changes in the years following the technical developments, the traffic and driving conditions.

Chapter 6

Robustness of filter and smoother distributions

In this chapter, we examine how sensitive the density $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$ is to the presence of extreme values in the observation series $(\widetilde{y}_t)$. This is useful to understand how robust the developed filtering and smoothing recursions are in the presence of outliers.

The discussion is inspired by the results by Dawid (1973) and O'Hagan (1979). In both cases, the inference of a location parameter given a random sample was considered from the Bayesian point of view. Precisely, Dawid (1973) considered a single observation x on a location model $X = \theta + D$, where θ was the unknown parameter with specified prior distribution P_1 and D was the unknown error with known error distribution P_2. He gave sufficient regularity conditions on the tails of P_1 and P_2 to imply that, as x tended to infinity, the posterior distribution of θ approached P_1 (x was discredited as an outlier), and similarly for the posterior mean of a function of θ. In addition, he observed that, since θ and D entered symmetrically, the same conditions on P_2 and P_1 implied that the posterior distribution of D approached P_2 (the prior was discredited; θ was *asymptotically fiducially* distributed). In particular, normal P_1 and student P_2 gave the former case and the reverse, the latter. O'Hagan (1979) proved that any admissible inference procedure applied to a t sample effectively ignored extreme outlying observations regardless of the assumed prior information. On the other hand, he showed that the normal distribution did not exhibit the same

behaviour. In fact, it never allowed outlier rejection, regardless of the prior information.
We prove now some similar results for our model using the L_1-norm. Without loss of generality, we assume that all observations $Y_{1:T}$ and all external regressors $(X_{t,1},\ldots,X_{t,m})_{(t=1:T)}$ are known. In addition, we write the model as in (2.6) and (2.7) such that outliers in the external regressors are also taken into consideration.

First, it is useful to summarize the definition and some well-known equivalent forms of the L_1-norm for densities:

$$\begin{aligned}\|f-g\|_1 &:= \int |f(x)-g(x)|\,d\mu(x)\\ &= 2\int (f(x)-g(x))^+\,d\mu(x) &(6.1)\\ &= 2\left(1-\int \min(f,g)\ d\mu(x)\right) &(6.2)\\ &= 2\ \sup_A |F(A)-G(A)| &(6.3)\\ &= 2\ \sup_{\Delta(\psi)\le 1}\left|\int \psi(x)\ \ (f(x)-g(x))\,d\mu(x)\right|.\end{aligned}$$

Here $x^+ := \max(x,0)$ is the positive part and $\Delta(\psi)=\sup_{x,x'}|\psi(x)-\psi(x')|$ is the variation of ψ. The equality (6.1) is Scheffé's theorem, see Devroye (1987), p. 2. The other equalities are easy consequences of this result. According to (6.3), the L_1-norm for densities is twice the total variation norm between the corresponding distributions.

Second, we prove an auxiliary result. It gives a lower and an upper bound for the ratio of the observation densities with different values of the state variable Z_t.

Lemma 6.1 *Let z_t and z'_t be given and let $\widetilde{Y}_t|Z_t$ have density $p_{VII}\left(m,c,z_t\right)(\widetilde{y}_t)$. Then, for all $\widetilde{y}_t$,*

$$\left[3+\left(\frac{z_t-z'_t}{c}\right)^2\right]^{-m} \le \frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|z'_t\right)} \le \left[3+\left(\frac{z_t-z'_t}{c}\right)^2\right]^{m}.$$

Proof: To begin, we rewrite the problem in an easier way using the definition of the Pearson type VII density. We have

$$\frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|z_t'\right)}=\frac{p_{VII}\left(m,c,z_t\right)\left(\widetilde{y}_t\right)}{p_{VII}\left(m,c,z_t'\right)\left(\widetilde{y}_t\right)}=\left[\frac{1+\left(\frac{\widetilde{y}_t-z_t'}{c}\right)^2}{1+\left(\frac{\widetilde{y}_t-z_t}{c}\right)^2}\right]^m=\left[\frac{1+(y+z)^2}{1+y^2}\right]^m$$

with $y:=\frac{\widetilde{y}_t-z_t}{c}$ and $z:=\frac{z_t-z_t'}{c}$. Thus, we have that

$$\frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|z_t'\right)}\geq\min_{\widetilde{y}_t}\frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|z_t'\right)}=\left[\min_y\frac{1+(y+z)^2}{1+y^2}\right]^m$$

and analogously

$$\frac{p\left(\widetilde{y}_t|z_t\right)}{p\left(\widetilde{y}_t|z_t'\right)}\leq\left[\max_y\frac{1+(y+z)^2}{1+y^2}\right]^m.$$

Therefore, we can reduce the attention to the function

$$h_z(y):=\frac{1+(y+z)^2}{1+y^2}.$$

The next step consists of finding a lower and an upper bound for it. To this end, we find its minimum and maximum. They should fulfill the condition

$$\begin{aligned}0=\frac{d}{dy}h_z(y)&=\frac{2(y+z)(1+y^2)-\left[1+(y+z)^2\right]2y}{(1+y^2)^2}\\&=\frac{2y+2y^3+2z+2zy^2-2y-2y^3-4zy^2-2yz^2}{(1+y^2)^2}\\&=\frac{-2z(y^2+zy-1)}{(1+y^2)^2}.\end{aligned}$$

The case $z=0$ means that $z_t=z_t'$ and then $\frac{p(\widetilde{y}_t|z_t)}{p(\widetilde{y}_t|z_t')}=1$. The assertion of the lemma follows trivially.

For the case $z\neq 0$, the two roots $y_\pm=\frac{-z\pm\sqrt{z^2+4}}{2}$ are always real and different. We assume for the moment that z is positive. Then, it follows from the sign of $\frac{d}{dy}h_z(y)$ that y_- is a minimum and y_+ is a maximum of the function $h_z(y)$. Consequently,

$$\begin{aligned}\max_y h_z(y)&=h_z(y_+)\\&=\frac{1+\left(\frac{z}{2}+\frac{\sqrt{z^2+4}}{2}\right)^2}{1+\left(-\frac{z}{2}+\frac{\sqrt{z^2+4}}{2}\right)^2}\end{aligned}$$

$$\begin{aligned}
&\leq 1 + \frac{1}{4}\left(z^2 + z^2 + 4 + 2z\sqrt{z^2+4}\right) \\
&\leq 1 + \frac{1}{4}\left(2z^2 + 4 + 2z\sqrt{\left(z + \frac{2}{z}\right)^2}\right) \\
&= 1 + \frac{1}{4}\left(2z^2 + 4 + 2z\left|z + \frac{2}{z}\right|\right) \\
&= 1 + \frac{1}{4}(4z^2 + 8) = 3 + z^2.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\min_y h_z(y) = \min_y h_z(-y-z) = \min_y \frac{1+y^2}{1+(y+z)^2} = \min_y \frac{1}{h_z(y)} \\
= \frac{1}{\max_y h_z(y)} \geq \frac{1}{3+z^2}.
\end{aligned}$$

Thus, we have proved the lemma for positive z. The case with negative z can be reduced to the positive one since

$$h_z(-y) = \frac{1+(-y+z)^2}{1+(-y)^2} = \frac{1+(y+|z|)^2}{1+y^2} = h_{|z|}(y).$$

□

We now examine what happens if some of the observations $(\widetilde{y}_t)$ go to ±∞. We show that these observations are discredited, i.e. they do not affect the resulting density in the limit.

Lemma 6.2 *Let I be any subset of $\{1,\ldots,T\}$. Then*

$$\lim_{\{\widetilde{y}_t, t\in I\}\to\pm\infty} \left\| p\left(z_{1:T}|\widetilde{y}_{1:T}\right) - p\left(z_{1:T}|\{\widetilde{y}_t, t\notin I\}\right)\right\|_1 = 0.$$

Remark 6.1 *The notation $\{\widetilde{y}_t, t \in I\} \to \pm\infty$ means that each $\widetilde{y}_t$, $t \in I$, goes to ∞ or $-\infty$ independently of the others.*

Proof: The proof follows by applying Lemma 6.1 and twice the Lebesgue's dominated convergence theorem.
The first goal is to compute $\lim_{\{\widetilde{y}_t, t\in I\}\to\pm\infty} p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$. We have

$$p\left(z_{1:T}|\widetilde{y}_{1:T}\right) = \frac{p\left(z_{1:T}\right)\ p\left(\widetilde{y}_{1:T}|z_{1:T}\right)}{p\left(\widetilde{y}_{1:T}\right)}$$

$$= \frac{p\left(z_{1:T}\right)\prod_{t\notin I} p\left(\widetilde{y}_t|z_t\right)\prod_{t\in I} p\left(\widetilde{y}_t|z_t\right)}{\int\cdots\int p\left(z'_{1:T}\right)\prod_{t\notin I} p\left(\widetilde{y}_t|z'_t\right)\prod_{t\in I} p\left(\widetilde{y}_t|z'_t\right)dz'_1\ldots dz'_T}$$
$$= \frac{p\left(z_{1:T}\right)\prod_{t\notin I} p\left(\widetilde{y}_t|z_t\right)}{\int\cdots\int p\left(z'_{1:T}\right)\prod_{t\notin I} p\left(\widetilde{y}_t|z'_t\right)\prod_{t\in I}\frac{p(\widetilde{y}_t|z'_t)}{p(\widetilde{y}_t|z_t)}dz'_1\ldots dz'_T}.$$

The terms which depend on $\{\widetilde{y}_t, t\in I\}$ are now only in the denominator. We succeed in computing the above mentioned limit if we apply Lebesgue's theorem to the denominator. To this aim, we need a majorant of the integrand which is both independent of $\{\widetilde{y}_t, t\in I\}$ and integrable with respect to $z'_{1:T}$. Moreover, we should compute the limit of the integrand.
The integrand can be bounded using Lemma 6.1:

$$p\left(z'_{1:T}\right)\prod_{t\notin I}p\left(\widetilde{y}_t|z'_t\right)\prod_{t\in I}\frac{p\left(\widetilde{y}_t|z'_t\right)}{p\left(\widetilde{y}_t|z_t\right)} \le p\left(z'_{1:T}\right)\prod_{t\notin I}p\left(\widetilde{y}_t|z'_t\right)\prod_{t\in I}\left(3+\left(\frac{z'_t-z_t}{c}\right)^2\right)^m.$$

The right hand side does not depend on $\{\widetilde{y}_t, t\in I\}$ and thus it can be chosen as majorant. We need the following inequalities to prove that it is integrable with respect to $z'_{1:T}$:

- For any functions $f_1, \ldots, f_n : \mathbb{R}^l \longrightarrow \mathbb{R}$, we have
$$\int\cdots\int\left|\prod_{i=1}^n f_i(z_{1:l})\right| dz_1\ldots dz_l \le \prod_{i=1}^n\left(\int\cdots\int\left|f_i(z_{1:l})\right|^n dz_1\ldots dz_l\right)^{1/n}$$
(generalized Hölder inequality).
- For any $a, b\in\mathbb{R}^+$, $c, d\in\mathbb{R}$ and $n\in\mathbb{N}$, we have
$$(a+b)^n \le (2a)^n + (2b)^n,$$
$$(c-d)^2 \le 2c^2 + 2d^2.$$
- For any z_t and $\widetilde{y}_t$, we have
$$p\left(\widetilde{y}_t|z_t\right) = b_{m,c}\left(\widetilde{y}_t - z_t\right) \le b_{m,c}\left(0\right)$$
(see Chapter 2 for the notation).

Thus,

$$\int\cdots\int p\left(z'_{1:T}\right)\prod_{t\notin I}p\left(\widetilde{y}_t|z'_t\right)\prod_{t\in I}\left(3+\left(\frac{z'_t-z_t}{c}\right)^2\right)^m dz'_1\ldots dz'_T$$
$$\le b_{m,c}\left(0\right)^{T-|I|}\ \mathbf{E}_{Z'_{1:T}}\left[\prod_{t\in I}\left(3+\left(\frac{z'_t-z_t}{c}\right)^2\right)^m\right]$$

$$
\begin{aligned}
&\leq b_{m,c}\left(0\right)^{T-|I|} \prod_{t\in I} \mathbf{E}_{Z_t'}\left[\left(3+\left(\frac{z_t'-z_t}{c}\right)^2\right)^{m|I|}\right]^{1/|I|} \\
&\leq b_{m,c}\left(0\right)^{T-|I|} \prod_{t\in I} \left\{6^{m|I|}+\left(\frac{2}{c^2}\right)^{m|I|}\mathbf{E}_{Z_t'}\left[(z_t'-z_t)^{2m|I|}\right]\right\}^{1/|I|} \\
&\leq b_{m,c}\left(0\right)^{T-|I|} \prod_{t\in I} \left\{6^{m|I|}+\left(\frac{2}{c^2}\right)^{m|I|}\mathbf{E}_{Z_t'}\left[\left(2(z_t')^2+2z_t^2\right)^{m|I|}\right]\right\}^{1/|I|} \\
&\leq b_{m,c}\left(0\right)^{T-|I|} \prod_{t\in I}\left\{6^{m|I|}+\left(\frac{8z_t^2}{c^2}\right)^{m|I|}+\left(\frac{8}{c^2}\right)^{m|I|}\mathbf{E}_{Z_t'}\left[(z_t')^{2m|I|}\right]\right\}^{1/|I|}.
\end{aligned}
$$

The last expression is finite since Z_t' is normally distributed, and therefore all moments exist and are finite. In addition,

$$
\begin{aligned}
\lim_{\{\widetilde{y}_t,t\in I\}\to\pm\infty}\prod_{t\in I}\frac{p\left(\widetilde{y}_t|z_t'\right)}{p\left(\widetilde{y}_t|z_t\right)} &= \lim_{\{\widetilde{y}_t,t\in I\}\to\pm\infty}\prod_{t\in I}\left[\frac{1+\frac{1}{c^2}(\widetilde{y}_t-z_t)^2}{1+\frac{1}{c^2}(\widetilde{y}_t-z_t')^2}\right]^m \\
&= \lim_{\{\widetilde{y}_t,t\in I\}\to\pm\infty}\prod_{t\in I}\left[\frac{\widetilde{y}_t^{-2}+\frac{1}{c^2}(1-z_t\widetilde{y}_t^{-1})^2}{\widetilde{y}_t^{-2}+\frac{1}{c^2}(1-z_t'\widetilde{y}_t^{-1})^2}\right]^m \\
&= 1.
\end{aligned}
$$

We can now put the previous results together and apply Lebesgue's theorem to interchange the limit with the integral. It follows that

$$
\begin{aligned}
&\lim_{\{\widetilde{y}_t,t\in I\}\to\pm\infty} p\left(z_{1:T}|\widetilde{y}_{1:T}\right) = \\
&= \frac{p\left(z_{1:T}\right)\prod_{t\notin I}p\left(\widetilde{y}_t|z_t\right)}{\int\cdots\int \lim_{\{\widetilde{y}_t,t\in I\}\to\pm\infty}\left(p\left(z_{1:T}'\right)\prod_{t\notin I}p\left(\widetilde{y}_t|z_t'\right)\prod_{t\in I}\frac{p(\widetilde{y}_t|z_t')}{p(\widetilde{y}_t|z_t)}\right)dz_1'\dots dz_T'} \\
&= \frac{p\left(z_{1:T}\right)\prod_{t\notin I}p\left(\widetilde{y}_t|z_t\right)}{\int\cdots\int p\left(z_{1:T}'\right)\prod_{t\notin I}p\left(\widetilde{y}_t|z_t'\right)dz_1'\dots dz_T'} \\
&= \frac{p\left(z_{1:T},\{\widetilde{y}_t,t\notin I\}\right)}{\int\cdots\int p\left(z_{1:T}',\{\widetilde{y}_t,t\notin I\}\right)dz_1'\dots dz_T'} \\
&= p\left(z_{1:T}|\{\widetilde{y}_t,t\notin I\}\right).
\end{aligned}
$$

Thus, the lemma follows using the Proposition 2.29 in van der Vaart (1998). This latter proposition can be proved easily by applying

Fatou's or Lebesgue's theorem. □

Finally, we compute the supremum in the L_1-norm of the difference between two densities with some different observations. The supremum can be at most 2 which corresponds to the case where the two densities are completely separated. For the considered model (2.6) and (2.7), we show that the supremum is less than 2. Thus, the densities are not completely separated also in the worst situation, i.e. the presence of outliers does not cause a breakdown in the estimation of $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$. To prove the described characteristic, we should generalize a well-known inequality.

Lemma 6.3 (Chebyshev's inequality in $\mathbb{R}^n$)
Let $X = \left(X_1, \ldots, X_n\right)^{'}$ be a random vector in $\mathbb{R}^n$. Then, for any positive c,

$$P\left(\|X - \mathbf{E}\left[X\right]\|_2 \geq c\right) \;\leq\; \frac{1}{c^2}\sum_{i=1}^{n} Var\left(X_i\right).$$

Proof: The lemma follows easily by applying the generalized Chebyshev's inequality in one dimension to the random variable $\|X - \mathbf{E}\left[X\right]\|_2^2$:

$$P\left(\|X-\mathbf{E}\left[X\right]\|_2^2 \geq c^2\right) \leq \tfrac{1}{c^2}\mathbf{E}\left[\|X-\mathbf{E}\left[X\right]\|_2^2\right] = \tfrac{1}{c^2}\textstyle\sum_{i=1}^{n} Var\left(X_i\right). \quad \square$$

Lemma 6.4 *Consider any two observation series $\widetilde{y}_{1:T}$ and $\widetilde{y}^{\star}_{1:T}$. Then*

$$\sup_{\{\widetilde{y}_{1:T},\widetilde{y}^{\star}_{1:T}\}} \|p\left(z_{1:T}|\widetilde{y}_{1:T}\right) - p\left(z_{1:T}|\widetilde{y}^{\star}_{1:T}\right)\|_1 \;<\; 2.$$

Thus, it follows easily that

Corollary 6.1 *Consider any two observation series $\widetilde{y}_{1:T}$ and $\widetilde{y}^{\star}_{1:T}$. Let $I = \{t\ |\widetilde{y}_t \neq \widetilde{y}^{\star}_t\} \subseteq \{1, \ldots, T\}$. Then*

$$\sup_{\{(\widetilde{y}_t,\widetilde{y}^{\star}_t),t\in I\}} \|p\left(z_{1:T}|\widetilde{y}_{1:T}\right) - p\left(z_{1:T}|\widetilde{y}^{\star}_{1:T}\right)\|_1 \;<\; 2.$$

Proof of Lemma 6.4: The first step is to reformulate the assertion using (6.2). We find:

$$\begin{aligned}&\sup_{\{\widetilde{y}_{1:T},\widetilde{y}^{\star}_{1:T}\}} \|p\left(z_{1:T}|\widetilde{y}_{1:T}\right) - p\left(z_{1:T}|\widetilde{y}^{\star}_{1:T}\right)\|_1 \\ &\qquad\qquad \leq 2 - 2\int\cdots\int \inf_{\{\widetilde{y}_t,t\in\{1:T\}\}} p\left(z_{1:T}|\widetilde{y}_{1:T}\right) dz_1 \ldots dz_T.\end{aligned}$$

Then, the lemma follows if we show that

$$\int \cdots \int \inf_{\{\widetilde{y}_t, t\in\{1:T\}\}} p\left(z_{1:T}|\widetilde{y}_{1:T}\right) dz_1 \ldots dz_T > 0.$$

The main ideas to prove this inequality are the following ones. To begin, we write $p\left(z_{1:T}|\widetilde{y}_{1:T}\right)$ in the same way as in the proof of Lemma 6.2. It is convenient to write the involved observation density $p\left(\widetilde{y}_t|z_t\right)$ with the notation $b_{m,c}\left(\widetilde{y}_t - z_t\right)$ introduced in Chapter 2. Further, we find successively lower bounds for the integral. To this end, we use Lemma 6.1, then we reduce the integral on $\mathbb{R}^T$ to an integral on a suitable subset and we estimate it using Chebyshev's inequality (Lemma 6.3). In addition, we use that $(c-d)^2 \leq 2c^2 + 2d^2$ for any $c, d \in \mathbb{R}$. Explicitly,

$$\begin{aligned}
&\int \cdots \int \inf_{\{\widetilde{y}_t, t\in\{1:T\}\}} p\left(z_{1:T}|\widetilde{y}_{1:T}\right) dz_1 \ldots dz_T \\
&= \int \cdots \int \inf_{\{\widetilde{y}_t, t\in\{1:T\}\}} \frac{p\left(z_{1:T}\right)}{\mathbf{E}_{Z'_{1:T}}\left[\prod_{t=1}^{T} \frac{b_{m,c}(\widetilde{y}_t - z'_t)}{b_{m,c}(\widetilde{y}_t - z_t)}\right]} dz_1 \ldots dz_T \\
&\geq \int \cdots \int \frac{p\left(z_{1:T}\right)}{\mathbf{E}_{Z'_{1:T}}\left[\sup_{\{\widetilde{y}_t, t\in\{1:T\}\}} \prod_{t=1}^{T} \frac{b_{m,c}(\widetilde{y}_t - z'_t)}{b_{m,c}(\widetilde{y}_t - z_t)}\right]} dz_1 \ldots dz_T \\
&\geq \int \cdots \int \frac{p\left(z_{1:T}\right)}{\mathbf{E}_{Z'_{1:T}}\left[\prod_{t=1}^{T} \left(3 + \left(\frac{z'_t - z_t}{c}\right)^2\right)^m\right]} dz_1 \ldots dz_T \\
&\geq \int \cdots \int_{\left\{\frac{\|z_{1:T}\|_2}{\sqrt{T\ Var(Z_1)}} < d\right\}} \frac{p\left(z_{1:T}\right)}{\mathbf{E}_{Z'_{1:T}}\left[\prod_{t=1}^{T} \left(3 + \left(\frac{z'_t - z_t}{c}\right)^2\right)^m\right]} dz_1 \ldots dz_T \\
&\geq \frac{1}{\mathbf{E}_{Z'_{1:T}}\left[\prod_{t=1}^{T} \left(3 + \frac{2(z'_t)^2 + 2d^2 T\, Var(Z_1)}{c^2}\right)^m\right]} \cdot \int \cdots \int_{\left\{\frac{\|z_{1:T}\|_2}{\sqrt{T\, Var(Z_1)}} < d\right\}} p\left(z_{1:T}\right) dz_1 \ldots dz_T \\
&= \frac{1 - P\left(\|Z_{1:T}\|_2 \geq d\sqrt{T\ Var\left(Z_1\right)}\right)}{\mathbf{E}_{Z'_{1:T}}\left[\prod_{t=1}^{T} \left(3 + \frac{2(z'_t)^2 + 2d^2 T\ Var(Z_1)}{c^2}\right)^m\right]}
\end{aligned}$$

$$\geq \frac{1-d^{-2}}{\mathbf{E}_{Z'_{1:T}}\left[\prod_{t=1}^{T}\left(3+\frac{2(z'_t)^2+2d^2T\ Var(Z_1)}{c^2}\right)^m\right]} \tag{6.4}$$

$$> 0$$

for $d > 1$. In fact, the random vector $Z_{1:T}$ is multivariate $\mathcal{N}(0, \Sigma)$ distributed and therefore the denominator is finite (we can use an argumentation similar to the one in the proof of Lemma 6.2). In addition, the numerator does not vanish. Finally, we stress again that the inequality (6.4) follows by generalizing Chebyshev's inequality as in Lemma 6.3 and using that $\mathbf{E}\left[Z_t\right] = 0$ for all t and $Var\left(Z_1\right) = \cdots = Var\left(Z_T\right)$. $\square$

Chapter 7

Outlook

It is now time for some critical remarks on the developed algorithms and for the outlook.
As we have seen in Section 5.1, the estimates of the states given the parameters work really well. The developed methods can cope with outliers and the CPU times are reasonable. On the other hand, the Kalman state estimates become unreliable, already in the presence of mild outliers.
The developed Monte Carlo estimation methods also produce fine results for the models without outliers, with isolated outliers or mild patchy outliers, see Section 5.2. The developed methods encounter difficulties when the patchy outliers have a big magnitude. The reason is that the considered regression model assumes i.i.d. observation errors, see (2.1) and (2.2). Again, the method derived from the Kalman filtering recursion cannot cope well with outliers.

The main critical point of our developed algorithms is that the estimates are computed with the maximum likelihood method, i.e. using a numerical optimization algorithm. This could be problematic since optimization algorithms may end up in local minima (or maxima) although quite good starting values are provided. In addition, numerical optimization algorithms are not particularly fast. A promising idea to avoid the numeric optimization is to expand the state vectors to contain also the unknown parameters, see Kitagawa (1998). Then, particular care is needed to avoid too big setups to compute the state densities. In fact, this approach should also be fast to compute.

Other possible improvements to our methods are to extend them to cope well also with patchy outliers in the observations or with innovative outliers in the state errors. Again, particular care is needed to avoid too big setups.
The theory should also be improved, for example by analysing the breakdown point of the estimation methods.

As we see, the end of a thesis is actually the start point for a new one.

Bibliography

Akaike, H. (1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. *Ann. Inst. Statist. Math. 26*, 363–387.

Anderson, B. D. and J. B. Moore (1979). *Optimal filtering* (3 ed.). Englewood Cliffs, New Jersey: Prentice Hall, Inc.

Arya, S., D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu (1998). An optimal algorithm for approximate nearest neighbor searching in fixed dimensions. *Journal of the ACM 45*(6), 891–923.

Bickel, P. J., Y. Ritov, and T. Rydén (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics 26*(4), 1614–1635.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis; Forecasting and Control* (3 ed.). Englewood Cliffs, New Jersey: Prentice Hall, Inc.

Carlin, B. P., N. G. Polson, and D. S. Stoffer (1992). A Monte Carlo approach to nonnormal and nonlinear state-space modeling. *Journal of the American Statistical Association 87*, 493–500.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*(3), 541–553.

Chan, K. S. and J. Ledolter (1995). Monte Carlo EM estimation for time series models involving counts. *Journal of the American Statistical Association 90*(429), 242–252.

Cochrane, D. and G. Orcutt (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association 44*, 32–61.

Colberg, C. A., B. Tona, G. Catone, C. Sangiorgio, W. A. Stahel, P. Sturm, and J. Staehelin (2005). Statistical analysis of the vehicle pollutant emissions derived from several european road tunnel studies. *Atmospheric Environment 39*, 2499–2511.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika 60*, 664–667.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B: Methodological 39*(1), 1–38. With discussion.

Devroye, L. (1987). *A Course in Density Estimation*, Volume 14 of *Progress in Probability and Statistics.* Birkäuser, Basel.

Doucet, A. (1998, January). On sequential simulation-based methods for bayesian filtering. Technical Report 112, Department of Electrical Engineering, Cambridge University.

Doucet, A., N. de Freitas, and N. Gordon (Eds.) (2001). *Sequential Monte Carlo methods in practice.* Statistics for Engineering and Information Science. New York: Springer-Verlag.

Fox, A. J. (1972). Outliers in time series. *Journal of the Royal Statistical Society, Series B: Methodological 34*, 350–363.

Friedman, J. H., J. L. Bentley, and R. A. Finkel (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software 3*, 209–226.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis 15*(2), 183–202.

Geweke, J. (1991). Efficient simulation from the multivariate normal and Student-t distributions subject to linear constraints. In E. M. Keramidas (Ed.), *Computing Science and Statistics. Proceedings of the 23rd Symposium on the Interface*, pp. 571–578. Interface Foundation of North America.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings Part F: Communications, Radar and Signal Processing 140*, 107–113.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley Series in Probability and Math. Statistics. New York: Wiley.

Handschin, J. E. (1970). Monte Carlo techniques for prediction and filtering of non-linear stochastic processes. *Automatica–J. IFAC 6*, 555–563.

Handschin, J. E. and D. Q. Mayne (1969). Monte Carlo techniques to estimate the conditional expectation in multi-stage non-linear filtering. *Internat. J. Control (1) 9*, 547–559.

Hannan, E. J. and M. Deistler (1988). *The statistical theory of linear systems.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.

Harrison, P. J. and C. F. Stevens (1976). Bayesian forecasting. *Journal of the Royal Statistical Society, Series B: Methodological 38*(3), 205–247.

Hausberger, S., J. Rodler, P. Sturm, and M. Rexeis (2002). Emission factors for heavy duty vehicles and validation by tunnel measurements. *Atmospheric Environment 37*, 5237–5246.

HBEFA (1999). *Handbuch für Emissionsfaktoren des Strassenverkehrs (handbook of emission factors for road traffic).* Umweltbundesamt Berlin; Bundesamt für Umwelt, Wald und Landschaft Bern.

HBEFA (2004). *Handbuch für Emissionsfaktoren des Strassenverkehrs (handbook of emission factors for road traffic).* Umweltbundesamt Berlin; Umweltbundesamt Wien; Bundesamt für Umwelt, Wald und Landschaft Bern.

Hürzeler, M. (1998). *Statistical Methods for General State-Space Models.* Ph. D. thesis, Swiss Federal Institute of Technology Zurich.

Ibragimov, I. A. (1956). On the composition of unimodal distributions. *Theory of Probability and its Applications (Translation of Teorija Verojatnostei i ee Primenenija) 1*, 255–260.

Isard, M. and A. Blake (1996). Contour tracking by stochastic propagation of conditional density. In *ECCV '96: Proceedings of the 4th European Conference on Computer Vision-Volume I*, pp. 343–356. Springer-Verlag.

Jensen, J. L. and N. V. Petersen (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *The Annals of Statistics 27*(2), 514–535.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous Univariate Distributions* (2 ed.), Volume 2 of *Wiley Series in Probability and Mathematical Statistics.* NY: Wiley.

Jones, R. H. (1980). Maximum likelihood fitting of arma models to time series with missing observations. *Technometrics 22*, 389–395.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D. J. Basic Engrg. 82*, 35–45.

Kalman, R. E. and R. S. Bucy (1961). New results in linear filtering and prediction theory. *Trans. ASME Ser. D. J. Basic Engrg. 83*, 95–108.

Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association 82*(400), 1032–1063. With comments and a reply by the author.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics 5*(1), 1–25.

Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association 93*, 1203–1215.

Künsch, H. (2003, January). Recursive Monte Carlo filters: algorithms and theoretical analysis. Technical Report 112, Seminar für Statistik, ETH Zürich.

Künsch, H. R. (2001). State Space and Hidden Markov Models. In *Barndorff-Nielsen, Ole E. (ed.) et al., Complex stochastic systems. London: CRC Press. Monogr. Stat. Appl. Probab. 87, 109–173.*

Marsaglia, G. (1977). The squeeze method for generating gamma variates. *Comput. Math. Appl. 3*(4), 321–325.

Martin, R. D. and D. Thomson (1982). Robust-resistant spectrum estimation. In *Proc. IEEE, Volume 70*, pp. 1097–1115.

Masreliez, C. (1975). Approximate non-Gaussian filtering with linear state and observation relations. *IEEE Trans. Autom. Control 20*, 361–371.

McLachlan, G. J. and T. Krishnan (1997). *The EM algorithm and extensions*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. New York: John Wiley & Sons Inc. A Wiley-Interscience Publication.

O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference. *Journal of the Royal Statistical Society. Series B. Methodological 41*(3), 358–367.

Percival, D. B. and A. T. Walden (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*. Cambridge University Press.

Pitt, M. K. and N. Shephard (1999). Filtering via simulation: auxiliary particle filters. *Journal of the American Statistical Association 94*(446), 590–599.

R Development Core Team (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ripley, B. D. (1987). *Stochastic Simulation*. NY: Wiley.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods* (Second ed.). Springer Texts in Statistics. New York: Springer-Verlag.

Ruckdeschel, P. (2001). *Ansätze zur Robustifizierung des Kalman-Filters*. Number 64. Dissertation, Universität Bayreuth, Bayreuth.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*(1), 115–131.

Staehelin, J., C. Keller, W. Stahel, K. Schläpfer, T. Bürgin, U. Steinemann, and S. Schneider (1997). Modelling emission factors of road traffic from a tunnel study. *Environmetrics 8*, 219–239.

Staehelin, J., C. Keller, W. Stahel, K. Schläpfer, and S. Wunderli (1998). Emission factors from road traffic from a tunnel study (Gubrist tunnel, Switzerland). Part III: Results of organic compounds, SO_2 and speciation of organic exhaust emission. *Atmospheric Environment 32*, 999–1009.

Staehelin, J., K. Schläpfer, T. Bürgin, U. Steinemann, S. Schneider, D. Brunner, M. Bäumle, M. Meier, C. Zahner, S. Keiser, W. Stahel, and C. Keller (1995). Emission factors from road traffic from a tunnel study (Gubrist tunnel, Switzerland). Part I: Concept and first results. *Science of the Total Environment 169*, 141–147.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wei, G. C. G. and M. A. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*, 699–704.

Wei, W. W. S. (1990). *Time Series Analysis: Univariate and Multivariate Methods.* Redwood City: Addison-Wesley Publishing Co Inc.

Curriculum Vitae

1973	Born on 20 June in Mendrisio TI
1979 - 1984	Primary school in Balerna TI
1984 - 1988	Secondary school in Balerna
1988 - 1992	Liceo in Mendrisio
1992	Matura Typus C
1992 - 1998	Studies in Mathematics at the Swiss Federal Institute of Technology (ETH), Zürich
1998	Diploma in Mathematics (Dipl. Math. ETH)
1998 - 1999	Member of the statistical consulting group at the Seminar for Statistics (Department of Mathematics), ETH Zürich
1999 - 2002	Statistical consultant for the European project ARTEMIS (*A*ssessment and *R*eliability of *T*ransport *E*mission *M*odels and *I*nventory *S*ystems); PhD student and teaching assistant at the Seminar for Statistics, ETH Zürich
2002 - 2005	PhD student and teaching assistant at the Seminar for Statistics, ETH Zürich